ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Big Data Eco System for Data Security and Privacy

Bhavya.S, B.N.Veerappa

PG Scholar, Department of Computer Science and Engineering, UBDT College of Engineering, Davangere, India

Associate Professor, Department of Computer Science and Engineering, UBDT College of Engineering, Davangere, India

ABSTRACT: Security and privacy issues are magnified by the volume, variety, and velocity of Big Data. The diversity of data sources, formats, and data flows, combined with the streaming nature of data acquisition and high volume create unique security risks. Privacy and Security of Big Data is gaining high importance since recently all the technologies started to depend on Big Data. In this paper, we are going to discuss particularly Big Data and difficulty in maintaining the privacy and security of Big Data. The main goal is to propose a Big Data system that maintains the privacy and security of the information stored on the cloud.

KEYWORDS: AES encryption, data chenking, sybil.

I. INTRODUCTION

Recent technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics, and social networks, are making possible to capture, process, and share huge amounts of data – referred to as big data - and to extract useful knowledge, such as patterns, from this data and predict trends and events. Big data is making possible tasks that before were impossible, like preventing disease spreading and crime, personalizing healthcare, quickly identifying business opportunities, managing emergencies, protecting the homeland, and so on.

Big data is relevant for all components of our society. Industry is using big data for shifting business intelligence from reporting and decision support to prediction and next move decisions. This use of big data emphasizes that big data is critical for obtaining actionable knowledge. Governments are also interested in using big data and predictive analytics to improve decision making and transparency, to engage citizens in public affairs, to improve national security. Healthcare Represents another major area to which big data may offer novel opportunities. Learning health systems are currently focusing on turning health care data into knowledge, translating that knowledge into practice, and creating new data by means of advanced information technology.

While the open source framework has enabled the footprint of Hadoop to logically expand,[5] enterprise organizations face deployment and management challenges with big data. Hadoop's core specifications are still being developed by the Apache community and, thus far, do not adequately address enterprise requirements for robust security, policy enforcement, and regulatory compliance. While Hadoop may have its challenges, its approach, which allows for the distributed processing of large data sets across clusters of computers, represents the future of enterprise computing.

II. RELATED WORK

Ajit Gaddam et.al [2] details the security challenges when organizations start moving sensitive data to a Big Data repository like Hadoop. It identifies the different threat models and the security control framework to address and mitigate security risks due to the identified threat conditions and usage models. The framework outlined in this paper is also meant to be distribution agnostic.

Priya P. Sharma et.al [3] Hadoop projects treat Security as a top agenda item which in turn represents which is again classified as a critical item. Be it financial applications that are deemed sensitive, to healthcare initiatives, Hadoop is

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

traversing new territories which demand security-subtle environments. With the growing acceptance of Hadoop, there is an increasing trend to incorporate more and more enterprise security features. In due course of time, we have seen Hadoop gradually develop to label important issues pertaining to, what we summarize as 3ADE (authentication, authorization, auditing, and encryption) within a cluster. There is no dearth of Production environments that support Hadoop Clusters.In this paper, we aim at studying "Big Data" security at the environmental level, along with the probing of built-in protections and the Achilles heel of these systems, and also embarking on a journey to assess a few issues that we are dealing with today in procuring contemporary Big Data and proceeds to propose security solutions and commercially accessible techniques to address the same.

Vinod Sharma [4] introduces the big data technology along with its importance in the modern world and existing projects like hadoop which are effective and important in changing the concept of science into big science. Hadoop, Map Reduce and No SQL are the major big data technology. This paper also throws some light on other challenges and issues. The various challenges and issues in adapting and accepting Big data security and suggest some more security standards and concept that make robust hadoop ecosystem without any processing overhead.

III. PROPOSED SYSTEM

There are privacy and security risks for the Big Data Eco system as the Name Node and the Data Node have the entire control over the data. The user has no control over the data and so there is a need for establishing a trust between the user and the Name Node, i.e. we are authenticating the user, so that not everybody can access the data.

To make the system more secure, we need to implement randomized encryption techniques on the data. Map Reduce does the processing and implementation of the random encryption techniques. Map Reduce is going to break the encryption/decryption process and speed up the process so as to ensure that the performance or scalability of the system is not affected. Multiple encryption techniques are implemented with an assumption that if the hacker manages to compromise certain chunks of data, he will not be capable of gaining access to all of the data for misuse.

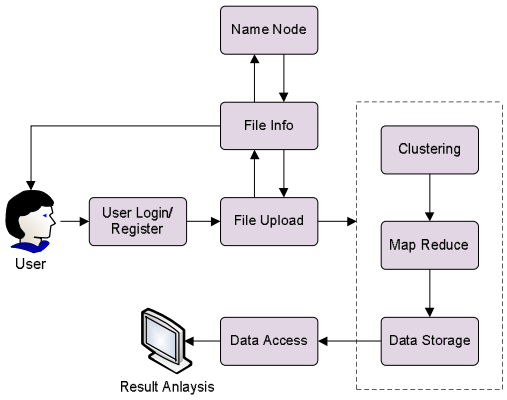

Figure1: Architecture of Proposed System

A. AES encryption

The AES algorithm operates on a 128-bit block of data and executed Nr - 1 loop times. The key length is 128, 192 or 256 bits in length respectively. The first and last rounds differ from other rounds in that there is an additional AddRoundKey transformation at the beginning of the first round and no MixCoulmns transformation is performed in the last round. In this paper, we use the key length of 128 bits (AES-128) as a model for general explanation. An outline of AES encryption is given in Fig. 1.a)

SubBytes Transformation: The SubBytes transformation is a non-linear byte substitution, operating on each of the state bytes independently. The SubBytes transformation is done using a once-precalculated substitution table called S-box.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

That S-box table contains 256 numbers (from 0 to 255) and their corresponding resulting values. This approach has the significant advantage of performing the S-box computation in a single clock cycle, thus reducing the latency and avoids complexity of hardware implementation.

ShiftRows Transformation: In ShiftRows transformation, the rows of the state are cyclically left shifted over different offsets. Row 0 is not shifted; row 1 is shifted one byte to the left; row 2 is shifted two bytes to the left and row 3 is shifted three bytes to the left.

MixColumns Transformation: In MixColumns transformation, the columns of the state are considered as polynomials over GF (28) and multiplied by modulo x4 + 1 with a fixed polynomial c(x), given by: c(x)={03}x3 + {01}x2 + {01}x + {02}.

AddRoundKey Transformation: In the AddRoundKey transformation, a Round Key is added to the State - resulted from the operation of the MixColumns transformation - by a simple bitwise XOR operation. The RoundKey of each round is derived from the main key using the KeyExpansion algorithm. The encryption/ decryption algorithm needs eleven 128-bit RoundKey, which are denoted RoundKey[0] RoundKey[10].

B. Data chenking

The basic element in the cloud storage system is chunk. A chunk is a data segment generated from a file. When user upload a file, if the file size is bigger than the configured size, it will be split into a collection of chunks. All chunks which are generated from a file except the last chunk have the same size (the last chunk of a file may have an equal or smaller size). After that, the ID generator will generate id for the file and the first chunk with auto-increment mechanism.

Next chunk in the chunks set will be assigned an ID gradually increase until the final chunk. A File Info object is created with information such as file-id, size of file, id of first chunk, number of chunks and stored. Similarly, the chunk will be stored in key-value store as a record with key is id of chunk and value is chunk data. Chunk storage is one of the most significant techniques. By using chunks to represent a file, we can easily build a distributed file storage system service with replication, load balancing, fault-tolerant and supporting recovery.

C. Dynamic workload Management

A dynamic load balancing algorithm makes load distribution decisions based on the current work load at every node of the distributed system. Accordingly, this algorithm must provide a means for collecting and managing system status information. The algorithm handles the requests in a proficient way. It starts by checking the counter variable of each server node and data center. After checking, it transfers the load accordingly by choosing the minimum value of the counter variable and the request is handled easily and takes a smaller amount of time, and offers maximum throughput. The randomly transfer of load can cause some server to heavily loaded while other server is lightly loaded. If the load is equally distributed it not only improves performance also reduces the time delay. This algorithm not only balances the load but also it improves the response time for the cloud.

While taking into account the impact of cost optimization one has to think on the subject of the solution to this difficulty. A counter variable is related with each node. Counter variable is the number of requests that the particular server node or data center is currently handling. Each node is having multiple data centers as shown in fig 3. The value of counter variable of server node will be equal to the sum of counter variables of its data centers. This algorithm essentially allocates request which is coming from the client nodes to the lightly loaded server cluster (Data Center) and gives the response in a reduced amount of time by doing this, it makes the algorithm proficient for response to request ratio. We can see that the clients at a same time make requests to access the cloud application over the internet

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

IV. RESULT AND DISCUSSION

Below figure shows file upload information.

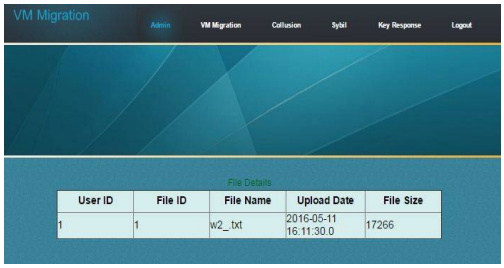

Figure 1: File Uploaded Information

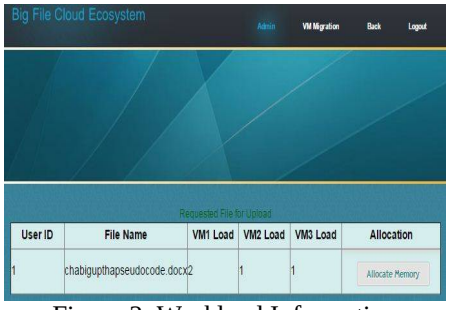

Figure 2: Workload Information

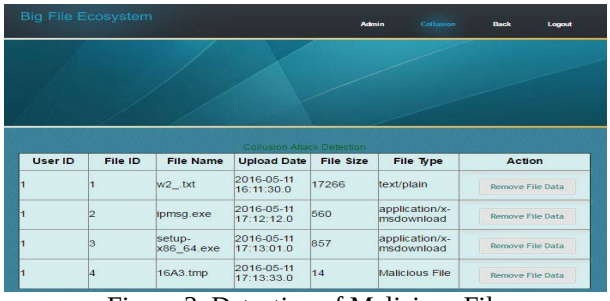

Figure 3: Detection of Malicious File

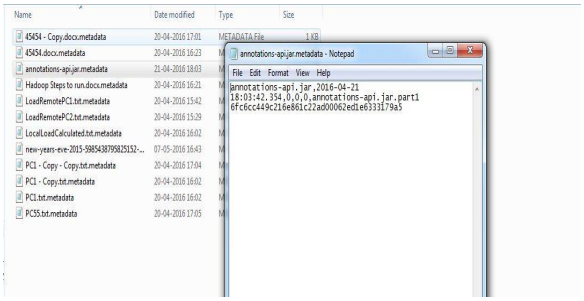

Figure 4:Meta Data Generation

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

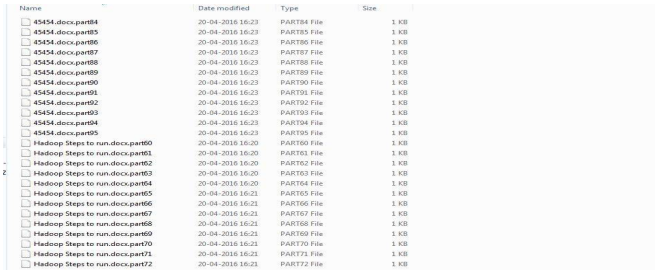

Figure 5: Data chucking

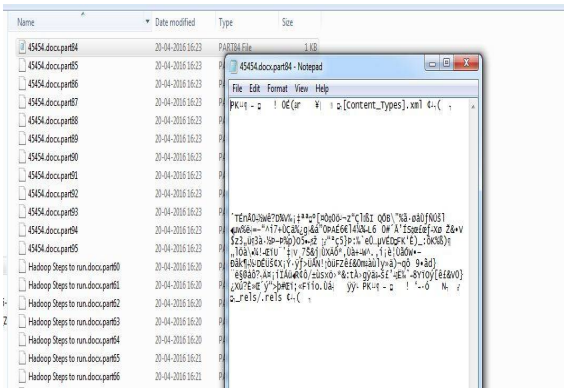

Figure 6: chunk information

V. CONCLUSION

By using our proposed system privacy and security of big data is achieved. Random encryption technique is used to store the data on cloud securely. Access to very large amount of private data is need to be taken care and securing the system is given with first priority.

REFERENCES

[1] Elisa Bertino, Big Data - Opportunities and Challenges. Cyber Center, CERIAS and CS Department. Computer Software and Applications Conference, IEEE 37th annual 2013.
[2] Ajit Gaddam, "Securing Your Big Data Environment",
[3] Priya P. Sharma, Chandrakant P. Navdeti, Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution. International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014.
[4] Vinod Sharma, Prof. N.K. Joshi," The Evolution of Big Data Security through Hadoop Incremental Security Model", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 5, May 2015.
[5] Nikita Haryani, Dhanamma Jagli," Dynamic Method for Load Balancing in Cloud Computing", IOSR Journal of Computer Engineering (IOSR-JCE). e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. IV (Jul – Aug. 2014), PP 23-28.
[6] Kalyani Shirudkar, Dilip Motwani," Big-Data Security", "International Journal of Advanced Research in Computer Science and Software Engineering", Volume 5, Issue 3, March 2015.