I.J. Information Engineering and Electronic Business, **2018, 2, 42-50**
Published Online March 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2018.02.06

Validation Lamina for Maintaining Confidentiality within the Hadoop

Raghvendra Kumar
Department of Computer Science and Engineering, LNCT College, Jabalpur, MP, India
Email: raghvendraagrawal7@gmail.com

Dac-Nhuong Le*
Faculty of Information Technology, Haiphong University, Haiphong, Vietnam
Email: Nhuongld@dhhp.edu.vn

Jyotir Moy Chatterjee
Department of Computer Science and Engineering, GD-RCET, India
Email: jyotirm4@gmail.com

Received: 07 June 2017; Accepted: 07 August 2017; Published: 08 March 2018

***Abstract*—**In the white paper we strive to cogitate vulnerabilities of one of the most popular big data technology tool Hadoop. The elephant technology is not a bundled one rather by product of the last five decades of technological evolution. The astronomical data today looks like a potential gold mine, but like a gold mine, we only have a little of gold and more of everything else. We can say Big Data is a trending technology but not a fancy one. It is needed for survival for system to exist & persist. Critical Analysis of historic data thus becomes very crucial to play in market with the competitors. Such a state of global organizations where data is going more and more important, illegal attempts are obvious and needed to be checked. Hadoop provides data local processing computation style in which we try to go towards data rather than moving data towards us. Thus, confidentiality of data should be monitored by authorities while sharing it within organization or with third parties so that it does not get leaked out by mistake by na ïve employees having access to it. We are proposing a technique of introducing Validation Lamina in Hadoop system that will review electronic signatures from an access control list of concerned authorities while sending & receiving confidential data in organization. If Validation gets failed, concerned authorities would be urgently intimated by the system and the request shall be automatically put on halt till required action is not taken for privacy governance by the authorities.

***Index Terms*—**Digital Signature, Electronic Signature, Data Local Processing, Hadoop, Big Data, Privacy Governance.

I. Introduction

The term Big Data was originally coined by John R. Mashey in late 1990. As a chief scientist at Silicon Graphics International (S.G.I.) in 1998 Mashey [1, 17, 21], he presented the issues of growing stress on storage & networking infrastructure in organizations that were growing quickly with data (image, graphics, models) & some more difficult data (audio, video). Then after Gartner officially phrased the term Big Data during his research in enterprise market. Gartner analyzed databases of commercial organizations and found there are lot of enterprise level data that which is not structured and has never been used for large scale analysis. This kind of data he termed as "*Dark Data*". Significance of extracting intelligence out of this dark data was sensed at this stage. Trying making decision out of this intelligence is the heart of Big Data application area.

Today's era is facing a lot of complexity when it comes to data and the challenge is how companies can make sense of the correlation of all these different types of data in a competitive market with sustainability and in the most secured way. It is estimated that Big data will double every two years [5, 18]. Most of the companies are making ambitious investments to establish their monopoly by figuring out the secret to transform Big Data analysis to big insight, and insight into business opportunity. Thus, as big data analysis is getting more important to companies, the more important it will be to secure& govern the data in preserving reputation & legal repercussions. While dealing data at an overwhelming scale, it is impossible to think about data management in traditional ways [3]. When we want to start using Hadoop as technological tool, we would be actually executing a distributed system because from technological prospective Big Data is all about handling Volume, Velocity, Variety, Veracity of data & we need larger networks, storages and other resources to manage them. Challenges with distributed systems are basically node failures, bottleneck problems, data synchronization, coordination of nodes, distribution & fragmentation of the jobs. While there are lot of technologies that evolves around Big Data, Hadoop provide one stop solution in open source framework. In the early 2000s, some engineers at Google looked into the future and determined that while their current solutions for applications such as web crawling, query frequency, and so on were adequate for most existing requirements, they were inadequate for the complexity they anticipated as the web scaled to more and more users. These engineers determined that if work could be distributed across inexpensive computing nodes & interconnect them on a network in the form of a “cluster”, it would solve their problem [9, 22].

The current security provisions are not adequate enough & it creates technological hindrance in preserving confidential data like analysis reports of medical data, stock market data, customer data etc. The measure of vulnerability & potential risks can be estimated by the famous NSA Scandal. It is now believed to have irreversibly damaged data security method which once was considered as the most trusted system in the world. The world of big data when used with Hadoop offers a new set of challenges and obstacles that make security and governance a challenge because its implementation typically includes open source code. Thus, there are potential risks for unrecognized back doors and default credentials. Many individuals and organizations working with big data wrongly assume that they do not have to worry about security or governance. But the issue security cannot be procrastinated because the state of the art is constantly evolving. Hand in hand with the latest security strategies we need respective governance strategy too [4, 18, 19].

The combination of security and governance will ensure accountability by all parties involved in the information management & administration. Managing the security of information needs to be viewed as a shared responsibility across the organization. Simply implementing all the latest technical security controls is not going to guarantee risk prevention if your end users cannot have a clear understanding of their role in keeping all the data that they are working with regard to security. Thus, when the data from a variety of sources is introduced, variety of security risks also gets into the company and unintended consequences can endanger the company.

In an instance while collecting data from unstructured data sources such as social media sites we have to ensure that viruses or spam links are not buried in its content. If we introduce this data and make it part of our analytics system, we could be putting our company at risk. Maintaining a large number of keys can be impractical, and managing the storing, archiving, and accessing of the keys is difficult. Hadoop platform runs on commodity hardware which means the security concerns are of component level. Frameworks like Map-Reduce, Yarn, and Spark harnesses distributed computing between clusters of machines and execute user defined jobs across the nodes in the cluster. Thus, we don’t require external components like SAN and when Hadoop uses commodity hardware it treat them as parallel array of disks without taking data out of local system. The vulnerable attack surfaces are the nodes in a cluster that may not have been reviewed and inadequately firewalled servers. These days companies have to put a lot of trust in the people running the infrastructure and monitoring the accesses. This limitation put companies at risk of any undesired or mistaken activity.

For example, in the English county of Devon, a situation arose recently where a primary school inadvertently sent an email to parents, which leaked private data about 200 children – the information included, their date of birth, their educational needs and behavioral issues. Despite the sensitivity of the data, it was still shared across [6, 16, 24]. Our idea of introducing Validation Lamina comes here that if confidential data related to students would have been digitally signed by the concerned authority, the emails having the data generated by any machine would have been reviewed and automatically halted before sending it to all parents.

In a traditional data system normally, standard user is allowed to directly access data. Admin users takes care of admin privileges & generally does not possesses access to data, other way around too, standard user is not allowed to gain access over admin privilege. Any unauthorized user is not expected to have any of the access. However, during attacks, the authority is over stepped & security breaches are created anonymously. Taking the target case as Hadoop, we generally have some users with respective roles & privileges. These users access data with some model which are exclusive processes to carry out tasks assigned to them. Sometimes the source who puts the data into data storage are not obvious& needed to be checked.

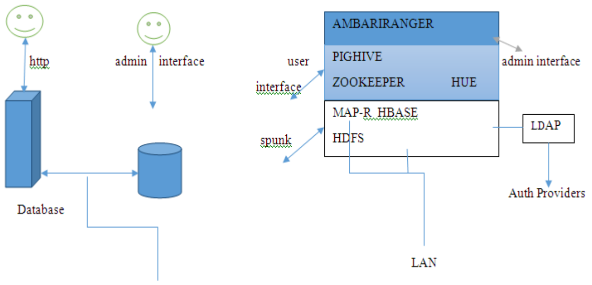

Fig.1. (Attack Points) Normal User System v/s Hadoop User System

Hadoop has its users as operator Ambari, Operator Ranger with user interface addons like Hive, Hue, Pig etc. as names related to animals conventionally. Dark data is typically fetched into Hadoop system for generating insights & creating monetization opportunities out of it. From attacker’s perspective, it is observed that within 20 minutes unauthorized penetration can be attempted successfully in a Hadoop system. Hadoop attack points can be divided among four major groups namely: User Interface, Admin Interface, and External Interface & Distribution Specific Interface.

Hue being classified as Internal Interface for attack point, uses Django. While installing Hue in Hadoop, DOM properties can be tempered by sending XSS to authorized Hadoop user& just an alert would be generated during the payload. Now, the sender of the XSS will have access to data as the Hadoop user do. Default configurations which are providing the distribution for plenty of applications in Hadoop also exposes the vulnerable information. Apache Ambari too act as a low hanging fruits to attacker by showing directory listing by default with no cookie flags & CSRF protection. Proxy script in here is also temper able for sending fake requests through exclusively accessible log file from DMZ (local network). Thus, exploitation of Hadoop servers in DMZ& access to logs become viable. Apache Ranger was also found to have missing function level access control& privileges can be escalated using CVE-2015-0266 to edit accounts, authorization rules and access policies. In new version, some of these vulnerabilities has been patched but CVE-2016-xxxx is still awaited. For distribution specific attack points, locally hosted as well as cloud related projects are targeted. Firms like Horton Works & Cloud era provides both ways hosting provisions. External Interface attack points includes internally running applications, vendors, monitoring modules and Auth providers like LDAP, Kerberos etc. Mostly outdated software's running is observed to be acting as attack points. It is very difficult to keep all the tools updated in Hadoop system for Big Data applications & ongoing researches are intended to patch the issues.

II. Principle & Architecture of Digital Signature

A digital signature is a protocol that is implemented like a real signature which provides a unique flag for a sender, and enables others to identify a sender from that flag and thereby confirm an agreement. We would like to propose the concept of Validation Lamina where such an agreement is to be designed to work while sharing of data in Hadoop system. If it does not meet the agreement, sharing of data will be halted since it won't be able to open it without proper key.

Digital signatures possess unique properties of having:

1. Non-repudiation: It provides insurance if a message is sent by someone having digital signature gets identified.
2. Uniquely traceable source authenticity (from expected source only)
3. Inseparable from message
4. Immutable after being transmitted
5. Have recent one-time use and should not allow duplicate usage.

Although digital signature is traditionally used for signing a digital document but it can be also implemented for cryptography. Some use the terms cryptography and cryptology interchangeably in English, while others *(including US military practice generally)* use cryptography to refer specifically to the use and practice of cryptographic techniques and cryptology to refer to the combined study of cryptography and cryptanalysis [15, 16, 20, 24].

1. Authentication: It is analogous to signing a contract but in digital world.
2. Authorization: It is analogous to signing & identifying one's own signature in digital world.

Proof of authenticity of the issuer of the data is to be maintained under Hadoop roof.

The more important Big Data is yielding analytical results for companies, the more important it becomes to secure such data. Confidentiality & Security maintenance are complementing each other in Big Data environment. The requirement it posses has to be closely tailored for special business goals. We can describe some of the challenges as follows:

Data Access: Dark data or computed data which is accessed by user posses almost same level of technical configurations in Hadoop environment as non-Hadoop implementations. However, it is mandatory to have such accesses to be available only at the instance when authentic business people require it for inspection or communication. Traditional data storage frameworks offer variety of security provisions. If augmentation is done with a federated identity capability we can design a security system that would be providing genuine data access across several layers of Hadoop framework.

Application Access: Application Programming Interfaces (API) generally are designed to control data exchange between communicating bodies in a secured fashion. This offers adequate level of security for Hadoop implementation. In Big Data Analytics on cloud the implementation under Software-as-a-service is employed to provide common big data related services like accessing service generated Big Data, Data Analytics result etc directly to users in order to increase efficiency & reduce cost [8, 23].

Data Encryption: In traditional systems, exchange of Big Data exhausts the system's resources. It is very challenging to secure data by means of encryption as it consumes costly computing cycles. To encrypt only the data in need is a wiser approach & modularizing encrypting elements on various hierarchical level in Hadoop further simply the problem. According to a normal estimation the cost of storing for a year in a traditional system costs around $37000, $5000 for database appliance and only $2000 for Hadoop cluster.

The property of signature enables its usability for doing any number of authentications, with variety of methods separately or in combined. The electronic signature thus should be using fully featured & secure type of scheme to preserve the aspect. The signed documents through electronic method also rely upon Public Key Cryptography (PKC) to authenticate identity in general. It can be further encrypted to provide additional confidentiality if required. In earlier time, the digital signature scheme was comprised of simple passwords or digitized images of handwritten signatures.

This was neither relied on cryptography nor possessed computer readable characters.

Today in electronic world the encrypted data by electronic signature is checked for the alteration performing digest/hash algorithm on raw data. Based upon this PKC, PKI is used as an internationally accepted scheme for securing electronic interactions. It involves a pair of mathematically related keys with very large prime numbers of 1024 characters in length. A public-key infrastructure (PKI) is a set of hardware, software, people, policies, and procedures needed to create, manage, distribute, use, store, and revoke digital certificates [11, 22]. The term Public key is referred as it is distributed freely to any user whom public key owner wishes to interact securely. The Private key component of the PKC scheme is known only by the signer. It is used to sign the data that is aimed to be verified by public key only. To issue these private keys (electronic certificates) certificate authorities are responsible to act as central body to authenticate & maintain records of certificates & keys. Certificate Authority audit themselves & customers can be either enterprise operated or can be a trusted third party. Private keys have to be bought on the name of the user.

There are several other schemes of cryptographic algorithms. Based on the number of keys that are employed for encryption and decryption, and utilization by their application, we would list following schemes:

- *Secret Key Cryptography (SKC)*: Uses a single key for both encryption and decryption
- *Public Key Cryptography (PKC)*: Uses one key for encryption and another for decryption
- *Hash Functions*: Uses a mathematical transformation to irreversibly "*encrypt*" information

Secret Key Cryptography (SKC): With a secret or symmetric key algorithm, the key is a shared secret between two communicating parties. Encryption and decryption both use the same key. The Data Encryption Standard (DES) and the Advanced Encryption Standard (AES) are examples of symmetric key algorithms. Symmetric-key cryptography refers to encryption methods in which both the sender and receiver share the same key (or, less commonly, in which their keys are different, but related in an easily computable way). This was the only kind of encryption publicly known until June 1976 [10]. Secret (Symmetric) Key cryptography uses single key for encryption and decryption. There are two types of secret key algorithms:

- *Block ciphers*: In a block cipher, the actual encryption code works on a fixed-size block of data. This technique involves encryption of one block of text at a time. Decryption also takes one block of encryption text at a time. If the length of data is not on a block size boundary, it must be padded.
- *Stream ciphers*: Stream ciphers do not work on a block basis, but convert 1 bit (or 1 byte) of data at time. Stream ciphers technique involves the encryption of one plain text byte at a time. The decryption also happens one byte at a time.

Public (*Asymmetric*) Key Cryptography: Public key cryptography provides the same services as symmetric key cryptography in general, but it uses different keys for encryption and decryption. A key pair in a public key cryptography scheme consists of a private key and a public key. These keypairs are generated by a process that ensures the keys are uniquely paired with one another and that neither key can be determined from the other (Hale and Friedrichs,2000). Public key cryptography was conceived in 1976 by Diffe and Hellman [12] and in 1977, Rivest, Shamir and Adleman designed the RSA Cryptosystem [13].

Public-key cryptography is a cryptographic technique that enables users to securely communicate on an insecure public network, and reliably verify the identity of a user via digital signatures [14]. Public key cryptography is an asymmetric scheme that uses a pair of keys: a public key, which encrypts data, and a corresponding private key, or secret key for decryption. Each user has a key pair given to him. The public key is published to the world while the private key is kept secret. Anyone with a copy of the public key can then encrypt information that only the person having the corresponding private key can read. It is computationally infeasible to deduce the private key from the public key. Anyone who has a public key can encrypt information but cannot decrypt it. Only the person who has the corresponding private key can decrypt the information. Again here each entity in a public key system will be assigned a private key and a public key. Private keys are kept private, and public keys are published and accessible to anyone.

Hash Function: A one-way hash function takes variable-length input - say, a message of any length and produces a fixed-length output; say, 160-bits. The hash function ensures that, if the information is changed in any way - even by just one bit - an entirely different output value is produced. A cryptographic hash operation produces a fixed-length output string called a digest from a variable-length input string. For all practical purposes, the following statements are true of a good hash function:

- Collision resistant: If any portion of the data is modified, a different hash will be generated.
- One-way: The function is irreversible. That is, given a digest, it is not possible to find the data that produces it.

These properties make hash operations useful for authentication purposes. For example, you can keep a copy of a digest for the purpose of comparing it with a newly generated digest at a later date. If the digests are identical, the data has not been altered.

MD5: MD5 is a message digest algorithm developed by Rom Rivest. MD5 actually has its roots in a series of message digest algorithm, which were the predecessors of MD5, all developed by Rivest. The Original Message Algorithm was called MD. MD5 is quite fast and produces 128-bit message digests. Over the years, researchers have developed potential weaknesses in MD5. However, so far MD5 has been able to successfully defend itself against collisions. This may not be guaranteed for too long, though. After some initial processing; the input text is processed in 512-bit blocks (which are further divided into 16, 32-bit sub-blocks). The output of the algorithm is a set of four 32-bit blocks, which make up the 128-bit message digest.

SHA-1: SHA stands for Secure Hash Function. Sha-1 is a strong cryptographic hashing algorithm, stronger than MD5. Sha-1 is used to provide data integrity (it is a guarantee data has not been altered in transit) and authentication (to guarantee data came from the source it was supposed to come from). it was produced to be used with the digital signature standard. Sha-1 uses a 160-bit encryption key. It is cryptographically stronger and recommended when security needs are higher. Sha-1 has proven to be a strong hashing algorithm and no records of it being hacked so far. Sha-1 works with any input message that is less than 264 bits in lengths. Input text is processed in 512 bits blocks. The output of SHA is a message digest, which is 160 bits length (32 bits more than the message Digest produced by MD5).

In India, the storage of private keys is legalized by government to be used on FIPS USB e-tokens & has discontinued the scheme of storing it on web browsers. The cost of fees is also based on the level of security & liability limits are indicated in Table 1.

Table 1. Description of different classes

Class of DSC	Description
Class 1	These certificates shall be issued to individuals/private subscribers
Class 2	These certificates will be issued for both business personnel and private individuals use.
Class 3	These certificates will be issued to individuals as well as organizations

A standard called as X.509 defines the structure of digital certificate. One of the latest version of the standard is Version 3, called as X.509V3. X.509 version -3 is combination of version-1 of the X.509 and version-2 of the X.509.Version-3 of the X.509 standard has added many extensions to the structure of digital certificate (like key usage, certificate policy, policy mapping, authority key identifiers, subject key identifier etc.). We require an electronic signature that would encrypt Variety of Big Data and establish a shared key over an insecure channel so that only the authority having the signature key will be able to send the data successfully otherwise anyone just having access to data on its local machine would not be able to do so. The validation will be provided by Validation Lamina having Access Control List (ACL). Thus, cipher text of any data is produced by performing a hashing algorithm which will be reducing the data into a unique number called message digest [16, 17].

In Hadoop system (Figure 2) there are many standard users distributed across the network working under an agency. Identification of such standard users is crucial & we propose to do it using electronic signature for identity validation. In order to communicate with different administrative sections standard user must be trusted upon & use of electronic seal would do us the favor. In order to generate this process of trust standard users first should be supplied with electronic signature that is adaptable by Hadoop system. It is to be noted that here the classification of human & robotic standard users can be designed by making use of biometric devices. However, overheads would be too high to implement but also will aid to trace attacks effectively.

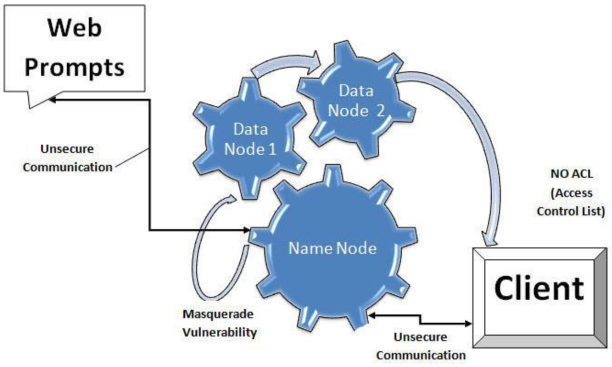

Fig.2. Hadoop’s Vulnerabilities

Our scheme is designed with aim to preserve performance and scalability while securing the Hadoop system. One must not consume computing cycles too much in doing encryption or it would slow down Big Data flow. Transparency is also a key factor for applications. No API changes should be made necessary so that legacy applications can continue to function. Further encryption can be used by administrator in a transparent manner.

III. ProposedScheme of Validation Lamina

The functional requirement of any Big Data technology is capturing data, next organizing it semantically and then integration. After when the first phase is accomplished, the data becomes compatible and ready for addressing various problems to be analyzed. Lastly management comes to action which is operationally automatic in Hadoop. Thus, architecture of any Big Data system must include variety of services on management level so as to make use of confidential data in the most secured & quick manner. Interfaces & feeds exist inside and outside of at all levels of architecture. It also tends to connect various levels by lying in between. We will have to apply our Lamina that should be acting as a firewall in between incoming feeds from external sources and outgoing data from internally managed resources. Next level will be redundant physical infrastructure which will store data in physical components having different locations. It has to be linked together through network that works on distributed computing model. Based on the choice of redundancy one would need to assign respective electronic signatures to all the sources for authenticating incoming data. This will make a provision to allow only selective authorities inside organization to access and administer the data. The data should be available for sharing only to those who are having the public key.

Although most of companies may go through an extensive background check on all of its employees, they have to keep a trust that no malicious insiders work in various business units outside of IT. They also have to assume that their cloud provider has diligently checked its employees. This concern is real because close to 50 percent of security breaches are caused by insiders (or by people getting help from insiders). Thus, Validation of documents will enable the tracking of any undesired activity related with the data being shared. Most core data storage platforms have rigorous security schemes that can be augmented with a federated identity capability through electronic signatures and providing appropriate access across the many layers of the architecture. In traditional environments, cryptography exhausts the system's resources completely. With the volume, velocity, and varieties associated with big data, this problem is exaggerated. Digital Signature Algorithm when perceived mathematically has its basis with discrete logarithm problem.

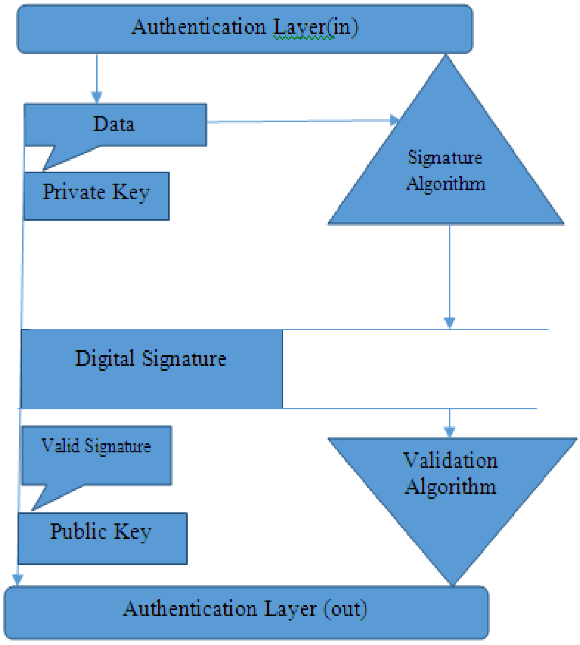

Fig.3. Architecture

The validation of Digital Certificate consists of the following steps:

i. The data digest algorithm calculates a message digest (hash) of all fields of the certificate, except for the last one, called MD1.
ii. The standard user passes all fields except the last one of the received digital certificate toa data digest algorithm. This algorithm should be the same as the one used by the CA while signing the certificate.
iii. The data digest algorithm calculates a message digest (hash) of all fields of the certificate, except for the last one, called MD1.
iv. The user now extracts the Digital certificate of the CA from the certificate. (it is a last field in a certificate).
v. The user de-Sign the CAs Signature (i.e. the user decrypts the signature with the CA s public key).
vi. This Produce another message digest, which we shall called MD2(i.e. MD2 is the same message digest as would have been calculated by the CA during the signing of the certificate.
vii. Now, the user compares the message digest it calculated with the one, which is the result of de-signing the CAs signature. if the two matches, i.e. if MD1=MD2, the user is convinced that the digital certificate was indeed signed by the CA with its private key. If this comparison fails, the user will not trust the certificate and reject it.

Although RSA encryption is faster than DSA but Hadoop environment gives us advantage of data local processing [7]. The only thing we require is faster key generation as data will be coming at enormous velocities through variety of sources in a Big Data system. Such a situation clearly idealizes the use of DSA over RSA since signature generation is faster in DSA. As compared to incoming data to a Big Data system outgoing data will be always small as we will be analyzing astronomical data into report level scale. Thus, even if signature validation is slower in DSA than RSA, overall performance of the system will be quicker. On priority, we need to make sure that our provider has the right controls in place to ensure that the integrity of the data is maintained. As such any modification is required to be assigned with a signature from legitimate authority. When Hadoop serve lets communicate with each other they do not verify other services what they really claims to be. Rogue can be easily stared therefore and Task-Tracker cans get access to data blocks. Hadoop also used web that serves interoperability among various nodes. Thus, we need to represent Digital-signature related information in a standard format called XML Signature.

The required data elements that are represented in an interoperable manner by XML Signature are:

1. Data to be digitally signed
2. Hash algorithm (MD5/SHA1) that will create the digest value
3. Signature algorithm
4. Semantics of certificate or key.

The logs containing list of access sessions (through sharing or forwarding) of confidential data with digital signature can be “seal” so that any change can be easily detected. This seal can be provided by computing a cryptographic function called hash or checksum, or a message digest. Thus hash function will be depending on all bits of the file being sealed and altering one bit will alter the checksum result. Each time the data is accessed or used, the hash function will be recomputed the checksum, and as long as the computed checksum matches the stored value, Validation can know if log of data has not been changed. The hash function and checksum confirm immutability of data hence signature will be required each time data is accessed by concerned authority. A Key-Store is to be applied at Validation Lamina which consist of a highly secured repository of keys or trusted certificates that are used for Validation, encryption etc. The entry of a key will be containing owner’s identity as private key. Trusted certificate’s entry will contain public key along with entity’s identity. For better management and security, we can use two Key-Stores with one containing the set of the key entries and other containing trusted certificate entries (including Certificate Authorities’ certificates). Again, access can be restricted to the Key-Store with private keys.

Trust-Store will contain certificates from a given list of expected authorities to communicate with or from Certificate Authorities that it trusts and identify other parties with proper protocol. JKS is most commonly used in the Java world and so can be ideally used as a Trust-Store in Hadoop. JKS doesn’t require each entry to be a private key entry, so it can be used for certificates from given trust but for which you don’t need Private keys.

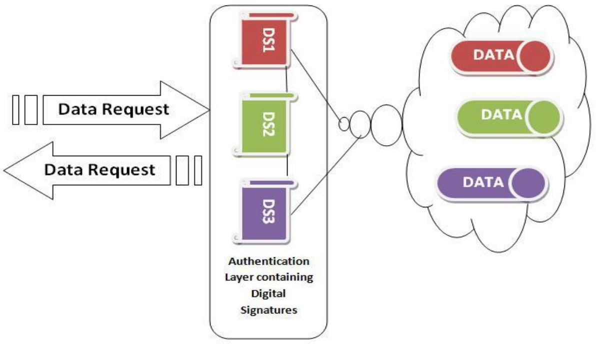

Fig.4. Validation Lamina

IV. Conclusion

In This article presented a framework for Validation while sharing of sensitive data on big data platform using electronic signature scheme. It will guarantee a faster & secured submission for accessing sensitive data that is based on encryption used by digital signatures algorithm among Hadoop servelets. In HDFS the basic idea of transparent encryption is, data is read & written into directories contained as sub tree of HDFS. All of the files in the sub tree are encrypted so that it helps applications be regulation complaint like HIPA, PCI DSS, FISMA etc. Encryption & decryption in the scheme is always done at client end and HDFS never handles unencrypted data or unencrypted keys. However, encryption can happen at any of several levels i.e. (Application, Database, File system, and Disk). HDFS encryption is somewhere between File system & Database level. On application level applying the encryption is most secure & flexible but probably not the best place as it is hardest to do. User developers need to rely on knowing the pitfalls & then develop the application in accordance. The reason it is hard to do is, it is hard to add to legacy applications. So, going back and adding own code becomes very difficult for developers. On Database layer, applying encrypting algorithms is pretty conventional. It incurs performance penalties plus inhibits lacks like secondary indices cannot generally be encrypted. Performance can be improved by implementing encryption on File System level where transparency is also preserved. There may be issues in terms of policies regarding safety where multiple users sitting on the top of a file system. Here we would be lacking flexibility since different encrypting policies for those users & tenets would have to follow the predefined scheme. Doing encryption at the Disk level would be resulting into a high performing system but would be only doing the protection against physical theft. It is desirable to have the encryption implemented as end to end i.e. getting it encrypted at application level and further doing it all the way down to the disk level. Data is thus prevented to be encrypted on network& preferred to get happen at-rest. Compartmentalization of data is also important and one of the course is key management being separated in orthogonal from HDFS management. Corollary is that HDFS admins & root users should not be allowed to access encrypted data unless they are not configured to do.

Electronic signatures being efficient in legally binding documents are complex to imitate and can be time-stamped. Thus, taking the best of two not only authenticates the data in & out of Hadoop Ecosystem but also creates cryptosystem to optimize the benefits of involved commodities under the semi-trusted conditions. Also, the security police in big market with analytics being performed on Hadoop framework will get facilitated by electronic signature scheme for tracing & locating the attack attempts quickly. The use of electronic signature improves performance without compromising at security fronts in Hadoop where data processing is expected to occurs in astronomical units. Future scope for the scheme is to design an expiration scheme to manage the generated signature. Keeping network access super tight by concealing configurations from internet is very much required to preserve the security. Also, user permissions should be minimally mapped for business roles unlike massively allowing all of the users. There is a need of a security counter measure that is independent of apps in Big Data environment, as Big Data is intended to handle Variety of known & unknown data source & structures. One must also aim to monitor & purge out obsolete software’s, external connections, bugs & CVE’s by listing all of the installed Hadoop components in a hierarchical manner. On HDFS level, all clients must be authenticated including tasks running as part of Map Reduce jobs & jobs submitted through Oozie. Users must also validate servers otherwise fraudulent servers could steal credentials. The idea of implementing validation lamina fulfills the requirement of prompting passwords during execution of a set of Hadoop commands & provides a single sign on feature. Otherwise trojan horse versions are easy to write. The three pillars of security are prevention, detection and response. The vast majority of the research in cryptography and protocols focuses on prevention. An area that has been largely ignored in the realm of private key compromise has to be researched further.

References

[1] Mashey, John R., 'Big Data and the Next Wave of Infra Stress', in SGI (1998).

[2] Davenport, Thomas H., Big Data in Big Companies Statistical Analysis System (SAS) Institute, 2013.

[3] S. Razick, R. Mocnik, L. F. Thomas, E. Ryeng, F. Drabløs, and P. Sætrom, The eGenVar data management system — Cataloguing and sharing sensitive data and metadata for the life sciences, Database, vol. 2014, p. bau027, 2014.

[4] Wang, Divyakant Agrawal-Amr El Abbadi-Vaibhav Arora-Ceren Budak-Theodore Georgiou-Hatem A. Mahmoud-Faisal Nawab-Cetin Sahin-Shiyuan, 'A Perspective on the Challenges of Big Data Management and Privacy Concerns', IEEE (2015).

[5] Gilder, Bret Swanson & George, 'Estimating the Exa flood', Tech rep., Discovery Institute, Seattle, Washington (2008).

[6] issuu.com, Big Data Innovation, Issue 12 by Innovation Enterprise.

[7] Bhushan Lakh, “Practical Hadoop Security”, Apress, 2014, Pg-151-154.

[8] Lyu, Zibin, Zheng-Jieming, Zhu-Michael R., ‘Service-generated Big Data and Big Data as a Service’, IEEE International Congress on Big Data (2013).

[9] Dean, Jeffrey and Ghemawat, Sanjay, ‘MapReduce: Simplified Data Processing on Large Clusters’, Google Inc, pp1-13 (2004).

[10] H. Zhu, D. Li, Research on Digital Signature in Electronic Commerce," The 2008IAENG International Conference on Internet Computing and Web Services, HongKong, 2008, pp. 807809.

[11] LPKI - A Lightweight Public Key Infrastructure for the Mobile Environments", Proceedings of the 11th IEEE International Conference on Communication Systems (IEEE ICCS'08), pp.162-166, Guangzhou, China, Nov. 2008.

[12] Diffe, W. and Hellman, M. E., New Directions in Cryptography. IEEE Transactions on Information Theory,22 (1 976), pp. 644-654.

[13] Rivest, R., Shamir, A. and Adleman, L., A Method for Obtaining Digital Signatures and Public Key Cryptosystems Communications of the ACM, 21(1978), pp.120-126.

[14] Adams, Carlisle & Lloyd, Steve (2003). Understanding PKI: concepts, standards, and deployment considerations. Addison-Wesley Professional. pp. 1115. ISBN 978-0-672-32391-1.

[15] Merriam-Webster's Collegiate Dictionary (11th ed.). Merriam-Webster.Retrieved2008-02-01.

[16] Lawrence, W., & Sankaranarayanan, S. (2012). Application of Biometric security in agent based hotel booking system-android environment. International Journal of Information Engineering and Electronic Business, 4(3), 64.

[17] Kaur, R. K., & Kaur, K. (2015). A New Technique for Detection and Prevention of Passive Attacks in Web Usage Mining. International Journal of Wireless and Microwave Technologies (IJWMT), 5(6), 53.

[18] Lasota, M., Deniziak, S., & Chrobot, A. (2016). An SDDS-based architecture for a real-time data store. International Journal of Information Engineering and Electronic Business, 8(1), 21.

[19] Nagesh, H. R., & Prabhu, S. (2017). High Performance Computation of Big Data: Performance Optimization Approach towards a Parallel Frequent Item Set Mining Algorithm for Transaction Data based on Hadoop MapReduce Framework. International Journal of Intelligent Systems and Applications, 9(1), 75.

[20] Kaur, P., & Monga, A. A. (2016). Managing Big Data: A Step towards Huge Data Security. International Journal of Wireless and Microwave Technologies (IJWMT), 6(2), 10.

[21] Maxwell, W. J. (2014). Global Privacy Governance: A comparison of regulatory models in the US and Europe, and the emergence of accountability as a global norm. Cahier de prospective, 63.

[22] Greene, D., & Shilton, K. (2017). Platform privacies: Governance, collaboration, and the different meanings of “privacy” in iOS and Android development. New Media & Society, 1461444817702397.

[23] Bhatti, H. J., & Rad, B. B. (2017). Databases in Cloud Computing: A Literature Review.

[24] Alguliyev, R. M., Gasimova, R. T., & Abbasli, R. N. (2017). The Obstacles in Big Data Process. *International Journal of Modern Education and Computer Science*, *9*(3), 28.

Authors’ Profiles

Raghvendra Kumar, Ph.D, is working as Assistant Professor in Computer Science and Engineering Department at L.N.C.T Group of College Jabalpur, M.P. India. He received B. Tech. in Computer Science and Engineering from SRM University Chennai (Tamil Nadu), India, M. Tech. in Computer Science and Engineering from KIIT University, Bhubaneswar, (Odisha) India and Ph.D. in Computer Science and Engineering from Jodhpur National University, Jodhpur (Rajasthan), India. He has published 86 research papers in international / National journal and conferences including IEEE, Springer and ACM as well as serve as session chair, Co-chair, Technical program Committee members in many international and national conferences and serve as guest editors in many special issues from reputed journals (Indexed By: Scopus, ESCI). He also received best paper award in IEEE Conference 2013 and Young Achiever Award-2016 by IEAE Association for his research work in the field of distributed database. His researches areas are Computer Networks, Data Mining, cloud computing and Secure Multiparty Computations, Theory of Computer Science and Design of Algorithms. He authored 12 computer science books in field of Data Mining, Robotics, Graph Theory, and Turing Machine by IGI Global Publication, USA, IOS Press Netherland, Lambert Publication, Scholar Press, Kataria Publication, Narosa, Edupedia Publication, S. Chand Publication and Laxmi Publication.

Dac-Nhuong Le has a M.Sc. and Ph.D in computer science from Vietnam National University, Vietnam in 2009 and 2015, respectively. He is Deputy-Head of Faculty of Information Technology, Haiphong University, Vietnam. Presently, he is also the Vice-Director of Information Technology Apply and Foreign Language Training Center in the same university**.** He has a total academic teaching experience of 12 years with many publications in reputed international conferences, journals and online book chapter contributions (Indexed By: SCI, SCIE, SSCI, Scopus, ACM, DBLP). His area of research includes: evaluation computing and approximate algorithms, network communication, security and vulnerability, network performance analysis and simulation, cloud computing, IoT and image processing in biomedical. His core work in network security, soft computing and IoT and image processing in biomedical. Recently, he has been the technique program committee, the technique reviews, the track chair for international conferences: FICTA 2014, CSI 2014, IC4SD 2015, ICICT 2015, INDIA 2015, IC3T 2015, INDIA 2016, FICTA 2016, ICDECT 2016, IUKM 2016, INDIA 2017, CISC 2017 under Springer-ASIC/LNAI Series. Presently, he is serving in the editorial board of international journals and he authored 4 computer science books by Springer, IGI Global, Lambert Publication, Scholar Press.

Jyotir Moy Chatterjee is working as Assistant Professor in Department of Computer Science and Engineering at GD-RCET, Bhilai, C.G, India. He received M. Tech from KIIT University, Bhubaneswar, Odisha and B. Tech in Computer Science & Engineering from Dr. MGR Educational & Research Institute University, Chennai, (Tamil Nadu). He is the member of CSI. His research interests include the cloud computing, big data, privacy preservation and data mining. He is also Oracle Certified OCA 10g and IBM Certified Associate System Administrator Lotus Notes and Domino 8.

How to cite this paper: Raghvendra Kumar, Dac-Nhuong Le, Jyotir Moy Chatterjee," Validation Lamina for Maintaining Confidentiality within the Hadoop", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.10, No.2, pp. 42-50, 2018. DOI: 10.5815/ijieeb.2018.02.06