Enhancing Big Data Security using Elliptic Curve Cryptography

Shubhi Gupta
Department of CSE
Amity University, Greater Noida
Uttar Pradesh, India
sgupta1@gn.amity.edu

Swati Vashisht
Department of CSE
Amity University, Greater Noida
Uttar Pradesh, India
svashisht@gn.amity.edu

Divya Singh
Department of CSE
Amity University, Greater Noida
Uttar Pradesh, India
dsingh@gn.amity.edu

Pradeep kushwaha
Department of CSE
Amity University, Greater Noida
Uttar Pradesh, India
pkkushwaha@gn.amity.edu

***Abstract*--Withgrowing times and technology, and the data related to it is increasing on daily basis and so is the daunting task to manage it. The present solution to this problem i.e our present databases, are not the long-term solutions. These data volumes need to be stored safely and retrieved safely to use. This paper presents an overview of security issues for big data. Big Data encompasses data configuration, distribution and analysis of the data that overcome the drawbacks of traditional data processing technology. Big data manages, stores and acquires data in a speedy and cost-effective manner with the help of tools, technologies and frameworks.**

Keywords - Big data, map reduce Hadoop, security and privacy, big data analytics.

I. INTRODUCTION

Big data is large data sets which are unable to be analyzed and managed by traditional processing systems. In big data, data sets grow to sizes which traditional IT's can no longer handle the size, scale and growth of data. The management and garnering value from it is difficult. The primary difficulties are the acquisition, storage, searching, sharing, analytics, and visualization of data. With evolving data set, the processes involved leveraging the data is also evolving. It is often synonymized with business intelligence, analytics and data mining. The difference between the two is that Big Data is about inductive statistics and business analytics is about descriptive statistics.

Big Data is not that something that has emerged in latest times but only in the last two years it saw an enormous amount of data recorded. Big Data has its cling to the fields of science and medicine, study of large and complex data has been done for drug development, physics modelling, and other forms of research. And now from these roots, Big data is starting to be evolved in different fields now.

Value extraction from the data set is easier than before. Big Data is full of challenges, ranging from the technical to the conceptual to the operational, any of which can derail the ability to discover value and leverage what Big Data is all about.Big data has its challenges ranging from technical to conceptual and operational which hampers the ability to better extract value and leverage the definition of big data.

II. LITERATURE SURVEY

As Big data is multi-dimensional, four primary aspects of it are:

1. Volume.
There is only one descriptive word for big data when it comes to its size; Large. Organizations can encompass terabytes and even petabytes of information.

2. Variety.
Not only structured data, Big data includes unstructured data as well. It could be anything from text, audio & video to click streams, log files, and more.

3. Veracity.
Huge amounts of data available for processing is prone to statistical errors and misinterpretation of the collected information. Purity plays a critical role.

4. Velocity.
As the data is sensitive, perishability of the data is an important concern. For value, it should be streamed as it is available and also archived.

These 4Vs of Big Data play a big role in laying out the path to analytics, with each having essential value in the process of unearthing value. But, not only these 4Vs make Big data as complex it is, there is another aspect to it which are: processes that Big Data drives.

Tools:

A. Hadoop

Hadoop is a tool used to deal with Big data for quite some time now. This has been there for a while, but now more and more different kind of businesses are leveraging and exploring its capabilities. The Hadoop platform cater to large structured and unstructured data sets in big data processing that does not suit to tables. It enables clustering and targeting. It supports analytics that are deep and computationally extensive.

Hadoop helps in managing the overheads associated with large data sets. In operation, when an organization's data are being loaded into a Hadoop software platform, it is broken down into manageable pieces and then automatically distributed to different servers. With this, it is ensured that there is no one place to go to access the data; the address where the data reside is tracked upon, and multiple copies of the data are created for safety. This leads to enhanced resiliency as one server goes missing in operation due to

some reason, that data can be replicated automatically with the help of a known copy of the same.

The Hadoop paradigm goes beyond than only working with data. for example, the traditional centralized database system, is limited to a large disk drive connected to a server class system featuring multiple processors. In this case, performance of the disk got hampered as the analytics is disturbed and, also the number of processors that can be conferred. With Hadoop clusters, each one of them participates in the processing of the data by spreading the work and the data across the cluster. each of the servers in the cluster in indexed with the jobs & then they operate upon themselves independently. The results are unified from each cluster and then delivered as a whole. This process, in Hadoop terminologies, is called MapReduce, where the processes, given codes, are mapped to all the servers in the clusters and the results are reduced to a singleton.

With this feature of hadoop, complex computational questions can be handled by harnessing all of the available cluster processors to work in parallel.

B. Map Reduce

A programming model or a software framework used in Apache Hadoop is Map Reduce. Hadoop MapReduce provides scalable, reliable and fault tolerant model where large data sets are processed and analyzed in parallel on large multinode clusters of commodity hardware. Analysis of data and processing is done in two different steps: Map phase and Reduce phase. A MapReduce acts first, breaks and divides the input data into chunks which is then first processed by Map phase and then by Reduce phase. The sorted output of the Map phase with the help of Hadoop, becomes an input to Reduce phase which initiates reduction parallelly. File system stores these files. MapReduce framework gets input datasets from HDFS22, 23 by default. Both the tasks might not be strictly sequential i.e., as soon as the map activity of an assigned set is complete, reduce activity can follow. All map activities might not be completed before any reduce activity happens. There is no such necessity. Both tasks of mapping and reducing works on key-value pairs. The input of the data set is taken as key-value pair and the processing of the output also generates in the form of key-value pair. Output from the Map phase is called intermediate results which becomes an input to reduce.

C. Cryptography & Public key Cryptography

Cryptography is a branch of applied mathematics that aims to add security in the ciphers of any kind of messages. Cryptography algorithms use encryption keys, which are the elements that turn a general encryption algorithm into a specific method of encryption. The data integrity aims to verify the validity of data contained in a given document[14]. A public key cryptosystem is an asymmetric cryptosystem where the key is constructed of a public key and a private key. The public key, known to all, can be used to encrypt messages. Only a person that has the corresponding private key can decrypt the message. The aim of this study is to analyze the performance and security of different public key cryptosystems over the fraudulence network for various applications such as image transmission, secure communication, E-messaging, large data transmission and etc. [7].

III. PROBLEM STATEMENT

Data, in today's time, carry standards pertaining to security governed by compliance laws and regulations. It could be of financial, medical, or government intelligence. It could be analytics set that needs protection. This data could be the same as what IT managers are coming across but Big Data analytics immingle the data and cross-index it which leads to the need of its security. IT managers should look for security solutions to the data stored in an array used for Big Data analysis. It should also be put under access authorization checks.

Privacy

Privacy and security concerns of the data gathered from the enterprises is not a new concept. However, Concept of Big data has done some benefit in this regard. Network personells do cater with a perimeter-based security mechanism such as firewalls but enforcements like these cannot prevent unauthorized access to data once a fraudulent has entered the network.

Challenges

Big data has been into the IT market for some time now but still problems are being faced in assembling the data and then analyzing it. Companies store different types of data differently(format). compiling, regularizing, and omission of irregularities without removing the information and its value is daunting and challenging.

Releasing information without authorization checks, changes in information and denial of services are examples of security breach. Now, this security can be achieved by proper authentication, preventing unauthorization, encryption and audit trials. Some of the techniques used are:

- Authentication method
- File encryption method
- Access control
- Key management
- Logging method
- Secure communication method

IV. SOLUTION

Algorithms in action:

One of the first published public-key algorithm was Diffie-Hellman. Computing discrete logarithmic has never been easy. This system however paved the way to compute exponents.In Diffie-Hellman, sender and receiver generates a secret key, this key is shared among the two in an insecure channel. They also share some information for computation the keys but still to know the key based on this information becomes difficult.

RSA public key cryptosystem shortly came after Diffie Hellman and is also one of the oldest and worked upon public key cryptosystems. This became first to sign as well

as encrypt. It works well with long keys and is widely accepted in Ecommerce applications.

Elliptic Curve Cryptography (ECC) proposed by Neal Koblitz and Victor Miller, has been in use for security reasons like key exchange and digital signatures. It works on graphical representation of coordinates which works for the calculation of the algorithm and a comparative level of security can be achieved with shorter keys.

NTRU works on the algebraic structures of polynomial rings. The main concern of the algorithm is to find a short vector in a given lattice. This reduces the polynomials with respect to two different moduli. It works in lesser time as RSA and ECC or any other public key system. As the computations are very simple, devices with restricted resources can also use this.

TABLEI.KEY SIZE RATIO OF ALL CRYPTOSYSTEMS

Diffie-Hellman Key size in bits	*RSA Key size in bits*	*NTRU Key size in bits*	*ECC Key size in bits*	*KEY SIZE RATIO(Bits)*
1024	1024	256	163	6:6:2:1
2048	2048	512	224	9:9:2:1
3072	3072	768	256	12:12:3:1
7680	7680	1920	384	20:20:5:1
15360	15360	3840	512	30:30:8:1

Key size of ECC comes out to be one sixth in reference to other cryptosystems and hence better.

V. PROPOSED ECC METHODOLOGY

We propose a secure cloud big data storage and its security using ECC algorithm. In the implementation, big data set is divided into sequential data parts based on same data type block or IP-resembled (Internet Protocol) data packets and is named alphanumerically.

ECC Cryptographic System [6]:

In this type of Public key cryptography, the user or the communicating device should have a pair of keys, public key and a private key. To carry the encryption and decryption process, some set of operations are performed on these keys. The underlying mathematic operation is defined over the elliptic curve y2 = (x3 + ax + b) mod p such that 4a3 + 27b2 mod p ≠ 0 where p is a large prime number and a and b are the coefficients that generates different elliptic curve points (x,y).

Operation:

1. Take a large prime no. p and values for coefficients a and b such that $4a^3 + 27b^2 \bmod p \neq 0$.
2. Consider an equation: $y^2 = (x^3 + ax + b) \bmod p$.
3. Take all values of y between 0 to p-1 and calculate y^2 mod p.
4. Take all values of x between 0 to p-1 and calculate $(x^3 + ax + b) \bmod p$.
5. Collect values of y from step 3 corresponding to values computed in step 4.
6. Collect all points (x,y) from step 5.
7. Input a supposed value of G known here as the base point which belongs to points from step 6.
8. Calculate 2G, 3G… such that:
   2G = G + G, 3G = 2G + G and so on until a value iG is found (let i be the least positive integer) such that the value of x coordinate of this point is same as the value of x coordinate of G and the value of y coordinate is prime number minus the value of y coordinate of G. From this, Order of G, called as n is computed as i+1.
   For instance, if p = 7, G = (1,3), values of 2G, 3G etc. will be calculated until the value of 3G comes out to be (1,4). Then the order, n will be 4.
   This addition will be done as follows:
   If point R = point P + point Q
   New point (x_R, y_R):
   $x_R = (\lambda^2 - x_P - x_Q) \bmod p$
   $y_R = ((\lambda (x_P - x_R) - y_P) \bmod p)$
   where $\lambda = (y_Q - y_P) / (x_Q - x_P) \bmod p$
   if P = Q and
   $(3x_P^2 + a) / 2y_P) \bmod p$ if P ≠ Q
9. Computation of sender's (say, A) public key:
   Choose a large integer n_A, so that it lies between 1 and n.
   Compute $P_A = n_A G$
10. Computation of receiver's (say, B) public key:
    Choose a large integer n_B, so that it lies between 1 and n.
    Compute $PB = n_B G$
11. Encryption:
    A will encrypt the message with B's public key –
    Let plain text P_m belongs to the point set in computed in step 6.
    Let k be the random integer which lies between 1 and n.
    Compute – $(kG, P_m + kP_B)$
12. Decryption:
    Compute kGn_B.
    Compute $P_m + k P_B - kGn_B$ to get P_m.

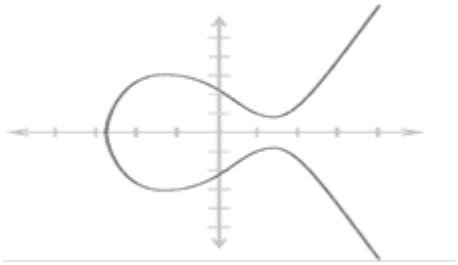

Fig. 1. An Elliptic curve.

This implements ECC using a data string which is an alphanumeric sequence. The encryption and decryption of a particular character is done by using its ASCII value.

For instance: a = 0,
b = -4,
Base point, G = (31, 6),
Sender's private key = 25,
Receiver's private key = 35,
Random key = 67.

This created over 200 different coordinates of an elliptic curve. The plaintext then becomes the coordinate stored at the number denoting the corresponding ASCII value.

- *Encryption:* A character is picked as plaintext. The ASCII value corresponding to it is taken into as an integer variable. The point on the elliptic curve corresponding to this particular integer is selected from the database. This following point is then encrypted. Now, this resultant point is mapped again to the database that will correspond to a new integer value. The new integer is then changed to a corresponding character which will consist of two specifications - printable ASCII character which further acts as an index and page number to which the corresponding index belongs to.
- *Decryption:* It selects the encrypted character and the coinciding page number. This calculates back the integer. Reverse mapping is carried out for the conversion of integer to point. Decryption is carried out. This again helps in getting integer from thedatabase. The corresponding character is the plain text character.
- The printable ASCII character ranges from 32 to 126 only. If the encryption character goes beyond this range, additional calculation is done. A tilde (~) is sent and the ASCII value gets incremented by 32 to send as a printable character whereas on the decryption side, reverse calculation is done when tilde is detected.

Plain Text: It works!!
Encrypted String: /~-~qI+RA##
Decrypted String: It works!!

VI. Result & Conclusion

The techniques of data handling in Big data were walk through-ed. Security of this data is an important factor to look into. A survey of several cryptographic techniques that can be used to secure data analytics were presented. Further research would result in more practical solutions to secure big data sets. The cost issues in terms of time and money need to be addressed. ECC algorithm does this efficiently by far and that too in a relatively less key size which make it easy to implement with less complexity. Data access becomes secure. Implementation and maintenance becomefairly easy. Reliability and scalability increase. Levels of services are guaranteed. Efficiency and security of the proposed scheme was concluded theoretically and also by comparing it to its peer algorithms. ECC is effective and feasible to protect the big data for cloud tenants.

References

[1] Sangita Bansal, Dr. Ajay Rana, Department of Computer Science and engineering Amity University, Noida (U. P.) India, transitioning from relational databases to big data, International journal of advanced research in computer science and software engineering valume4, Issue 1, January 2014

[2] Raghav Toshniwal, kanishka Ghosh Dastidar,Ashok Nath,department of computer science,st.xaviers college (autonomous) kollata,india , Big Data Security issue and challenge, International Journal of Innovative in Advanced Engineering (IJIRAE) ISSN:2349-2163 ISSUE 2, Volume 2 (February2015).

[3] Venkata Narasimha Inukolu, sailajaArsi and Srinivassa Rao Ravuri, Department of computer Engineering, texas tech university, USA Department of banking and financial services cognizant technology solution, India, International journal of network security and its application (IJNSA), vol 6 no. 3 May 2014

[4] Big_Data_Analytics_for_Security_Intelligence.pdf

[5] Vinit Gopal Savant,Department of computer engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India.,vinitsawant06@gmail.com, Approaches to Solve Big Data Security Issues and Comparative Study of Cryptographic Algorithms for Data Encryption, International Journal of Engineering Research and General Science Volume 3, Issue 3, May-June 2015, ISSN 2091-2730.

[6] Gupta Shubhi, Department of Computer Science and Engineering, Amity university, Greater Noida, "Implementation of ECC using socket programming in Java", International Organization of Scientific Research (IOSR), , Volume 16, Issue 4, Ver. I (Jul-Aug. 2014), PP 87-89

[7] Gupta Shubhi, Department of Computer Science and Engineering, Krishna engineering college, "Key based performance analysis of different public key cryptosystems: a survey", International Journal of Advanced research in computer science, , Volume 3,No.2

[8] William Stallings, Cryptography and network security, 2nd edition, Prentice Hall publications

[9] B.Schiener, Applied Cryptography. John Wiley publications and sons, 2nd edition, 1996

[10] Victor Miller, "Uses of elliptic curves in cryptography", Advances in cryptology, 1986

[11] N. Koblitz, A course in number theory and cryptography.

[12] http://www.tutorialspoint.com/java/java_networking.

[13] KohlekarMegha, Jadhav Anita, 2011." Implementation of Elliptic Curve Cryptography on Text and Image", International Journal of Enterprise Computing and Business Systems, Vol. 1 Issue 2 July 2011.

[14] Diego F. de Carvalho, Rafael Chies, Andre P. Freire, Luciana A. F. Martimiano and RudineiGoularte, "Video Steganography for Confidential Documents: Integrity, Privacy and Version Control", University of Sao Paulo – ICMC, Sao Carlos, SP, Brazil, State University of Maringa, Computing Department, Maringa, PR, Brazil.