Big Data Security
A Review of Encryption Techniques in Big Data

M. Geethanjali
Asst. Professor, Research Scholar,
Dept. of Computer Science,
St. Joseph's College of Arts and Science for Women,
Hosur.
geethaanjalisubramani@yahoo.com

Dr. P. Madhubala
HOD, Research Supervisor,
Don Bosco College ,
Dharmapuri.
madhubalasivaji@gmail.com

Abstract: In recent years, big data have been hot research topic. The interesting amount of big data also increases the chance of breaching the privacy of individuals. Due to a rapid growth and spread of network services, mobile devices, and online users on the internet leading to a remarkable increase in the amount of data. However, it is not only very difficult to store and analyse them with traditional applications. But also it has challenging data privacy and security problems. This paper shows the fundamental concept of big data, concerns on big data, technologies used and presents comparative view of big data privacy and security approaches in literature.

Keywords- Big data analytics, hadoop, security and privacy, encryption

1. INTRODUCTION

Big data refers to large volume of data in everyday lives. The data generation rate is growing rapidly. So, it is becoming extremely difficult to handle it using traditional methods or systems .The amount of data generated by social networking sites, sensor networks, internet, health care applications, and many other companies is drastically increasing day by day. All the huge amount of data generated from different source in multiple formats with very high speed is referred as big data. Mean while, all data generated may be structured form (relational data), semi-structured form (XML data) and un-structured form (document, pdf, image, video, audio, media logs, MRI scan report, X-rays, etc..) which is not managed by traditional databases (i.e., in rows and columns) (Divakar, Shrikant, & Shweta). Big data is heterogeneous data. Social media like face book, twitter, Google, generates huge amount of data daily, which is complex and unstructured in nature to handle by traditional databases.

Identifying the source of problems will result in more efficient use of big data. This paper examines and classifies studies on security and privacy breaches and solutions in big data. These perspectives would lead to an understanding of important research areas and the development of new methods. Security issues are the merging concepts in big data. (Miloslavskaya, Senatorov, Tolstoy, & Zapechnikov, 2014) In information security a lot of encryption algorithm is widely used. Asymmetric key known to be public key and symmetric key is known to be a secret key encryption is the two classifications of the encrypted algorithm.

2. IDENTIFICATION AND CHARACTERISTICS OF BIG DATA

Big data refers to large and complex data sets. The typical software is inadequate for managing. There are various explanations of big data Vs. 5Vs are typically used to characterize of big data as volume, velocity, variety, veracity and value. Figure 1 shows the big data.

1. Volume is the size of data
2. Velocity is the high speed of data
3. Variety indicates heterogeneous data types and sources
4. Veracity describes consistency and trust worthy of data
5. Value provides outputs for gains from large data sets

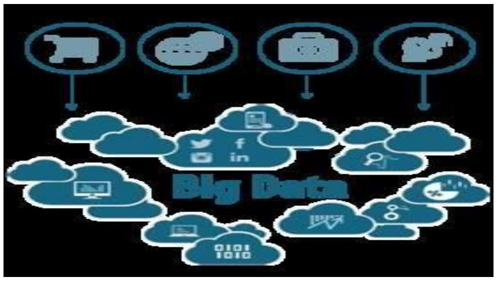

Figure 1. Big Data

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones and global positioning system (GPS) devices. Structured data also include things like account balances, sales figures and transaction data.Semi-structured data can contain both the forms of data. We can see semi-structured in form but it is actually not defined with a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.Unstructured data include more multifarious information, such as customer reviews from feasible websites, photos and other multimedia, and comments on social networking sites. These data cannot be separated into categorized or analyzed numerically.

3. CLASSIFICATION OF BIG DATA

Big data is classified into ten categories in terms of data type, data format, data usage, data analysis, data store, data frequency, data processing purpose method. Table 1 shows the classification of big data.

Table: 1 Classification of Big Data

Classification of Big data	Patterns
Data type	Transactional, Historical, Master, meta
Data Format	Structured, semi-structured, Unstructured
Data Source	Web and social media, Internet of things(IoT)
Data Consumer	Human, Business Process, Enterprise applications, Data Repositories
Data Usage	Industry , Academic, Research Centres
Analysis type	Interactive, Real time, Batched , Mix
Processing purpose	Predictive, Analytical, Modelling,
Processing method	High performance, computing, Distributed parallel, cluster, Grid
Data store	Relational, Graph, Key-value, Column oriented, Document oriented
Data frequency	On demand, Real Time, Time series

4. BIG DATA ANALYTICS

Big data analytics is the process of examining large sets of data containing a variety of data types in big data-to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information (paul, 2013). The analytical findings can offer a nearly endless source of business and information insight that can lead to operational improvement and new opportunities for companies to provide unrealized revenue across almost every industry (Miloslavskaya, Senatorov, Tolstoy, & Zapechnikov, 2014). The main goal of the big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modellers and other analytics professionals to analyse large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs social media content and social network activity reports, from customer emails and survey responses, mobile-phone call details records. Big data can include both structured and unstructured data. Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually for example, real-time data on the performance of mobile applications.

4.1. How Big Data Relates To Cloud Computing

In the cloud computing context, network accessible resources are defined as services. These services are typically delivered via one of three cloud computing service models:

a) Infrastructure as a services (Iaas) offers storage, computation, and network capabilities to service subscribers through virtual machines(VMs).

b) Pratform as a service (PaaS) provides an environment for software application development and hosts a client's applications in a PaaS provider's computing infrastructure.

c) Software as a service (SaaS) delivers on-demand software services via a computer network, eliminating the cost of purchasing and maintaining software.

Big Data analytics use computation intensive data mining algorithms that require efficient high performance processors to produce timely results.Cloud computing infrastructures can serve as an effective platform for addressing both the computational and data storage needs of big data already resides in the cloud, and this trend will increase in the future. This trends requires that clouds becomes the infrastructures for implementing pervasive and scalable data analytics platforms coping with and gaining value from cloud-based big data requires novel software tools and innovative analytics techniques. These technical and business advantages come at a cost. The security vulnerabilities inherited from the underlying technologies (that is, virtualization, IP, APIs and data center) prevent organizations from adopting the cloud in many critical business applications.

4.2. Cloud Based Data Analytics

Big Data refers to massive, heterogeneous, and often unstructured digital ontent that is difficult to process using traditional data management tools and techniques. The term encompasses the complexity and variety of data and data types, real-time data collection and processing needs, and the value that can be obtained by smart analytics advanced data mining techniques and associated tools can help extract information from large, complex data sets that is useful in making informed decisions in many business and scientific applications including tax payment collection, market sales, social studies, bio sciences and high energy physics (Saraladevi, Pazhaniraja, Paul, Basha, & Dhavachelvan, 2015). Combining big data analytics and knowledge discovery techniques with scalable computing systems will produce new insights in a shorter time. Although few cloud-based analytics platforms are available today, current research work anticipates that they will become common within a few years. Some current solutions are based on source systems such as Apache Hadoop and SciDB, while others are proprietary solutions provided by companies such as Google, IBM, EMC, BigML, Splunk strom.

5. BIG DATA TECHNOLOGIES TO ANALYSE DATA

- These big data technologies have the potential to dramatically re-invigorate your existing data warehouse and BI investments with new capabilities and new architectural approaches. Organizations have an opportunity to extend their existing data warehouse and BI environments by leveraging the following big data capabilities:
- Storage, access and analysis of massive volumes (meanings hundreds of terabytes and petabytes of data) of structured transactional data(such as sales, orders, shipments, point-of-sale transaction, call details records, and credit card transactions) at the lowest level of granularity
- Integration of semi structured data(for example, web logs, sensor, GPS, and telemetry data) and unstructured data(such as text fields, consumer comments, documents, and maintenance logs) that can add new dimensions, dimensional attributes, and metrics to your existing data warehouse and BI reports and dashboards
- Real-time data feeds coupled with real-time analytic environments for capturing, analyzing, flaging, and action on abnormalities in the data as it flows into your organization.
- Predictive analytics that can create scores, forecast, propensities, and recommendation that can be integrated into your key business operational systems(such as financial, call centres, sales, procurement, marketing, and other operational systems) and management systems (for example, alerts, reports, and dashboards.

5.1 Apache Hadoop

Apache Hadoop is an open-source software frame work that supports data-intensive, natively distributed, natively parallel applications. For many, hadoop has become synonymous with big data. It supports the running of applications on large clusters of commodity hardware using the scale-out architecture (Saraladevi, Pazhaniraja, Paul, Basha, & Dhavachelvan, 2015). Hadoop implements a computational paradigm named MapReduce where the applications is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In additional, hadoop provides a distributed filesystem(called the hadoop distributed file system, or HDFS) that stores data on the compute nodes, which provides very high-aggregate bandwidth across the cluster. Both MapReduce and HDFS are designed so that node failures are automatically handled by the framework. It enables application to work with thousands of computations-independent computers and petabytes of data. The entire Apache Hadoop "platform" is now commonly considered to consists of the hadoop kernel, MapReduce, HDFS, and a number of related projects including Apache Hive, Apache HBase, and others

5.2 Hadoop MapReduce

Map Reduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster (Xuyun, et al., 2013).A Map Reduce programme comprises a Map() produce that performs filtering and sorting(such as sorting students by first name into queues, one queue for each name) Reduce() procedure that performs summary operations(such as, counting the number of students in each queue, yielding name frequencies). The MapReduce system (also called infrastructure or framework) orchestrates the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, providing for redundancies and failures, and managing the overall process.

5.3 Apache Hive

Apache Hive is a data ware house infrastructure built on the top of hadoop that provides data summarization, query, and analysis. While initially developed by Facebook, Apache hive is nowd and is being enhanced by other companies, such a Netflix. Apache Hive supports analysis of large data sets stored in Hadoop- compatible file systems. It provides an SQL-like language called HiveQL while maintain full support for MapReduce. To accelerate queries, Hive provides indexes, including bitmap indexes.

5.4 Apache HBase

HBase is an open source, non-relational, distributed database model written in Java. It was developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS. HBase can serve as the input and output for Map Reduce jobs run in Hadoop.

5.5 Pig

Pig is a high-level, natively parallel data-flow language and execution frame work for creating MapReduce programs. Pig abstracts the Map Reduce programming language into a higher-level construct, similar to how SQL is a higher-level language for relational database management systems. Pig can be extended using user-defined functions, which the developer can write in Java, Python, JavaScript, or Ruby and call directly from the language.

6. NEW ANALYTIC ALGORITHM

- Support vector machines are based on the concept of decision planes that define decision boundaries and a decision plane that separates sets of objects having different class membership
- Random Forest consists of a collection of simple tree predictors, each capable of producing a response when presented with a set of predictor value.
- Ensemble methods are a model testing and verification technique that tests multiple models to obtain better predictive performance than could be obtained from any one analytic model. Text mining is a process to mine unstructured information and extract meaningful numerical metrics from the unstructured data, turning unstructured data into structured results.
- Feature selection is the process of selecting a subset of relevant features for use in model construction especially when the data may contain many redundant or irrelevant variables.

7. BIG DATA SECURITY TECHNIQUES

Organizations used various methods of identification to ensure security and privacy. The most common solution to ensure security and privacy may be oral or written pledges. Passwords, controlled access, and two factor authentication is low-level, but routinely used, technical solution to enforce security and privacy when sharing and aggregating data across dynamic, distributed data services. Access permissions such as these can potentially be broken by both the intentional sharing of permissions after they are no longer required or potential (Miloslavskaya, Senatorov, Tolstoy, & Zapechnikov, 2014).More advanced techniques solution is cryptography. The famous encryption schemes are AES and RSA. Virtual barriers such as firewalls, secure socket layer and transport layer security are designed to restrict access to data. Each of these technologies can be broken, however and thus need to be constantly monitored, with fixes applied as needed. Tracking, monitoring or auditing software is developed to provide a history of data flow and network access by an individual user in order to ensure compliance with security related.The limitation of this technology is that it is difficult and costly to implement on a large scale or with distributed systems and users because it requires dedicated staff to read and interpret the findings, and the software can be exploited to monitor individual behaviour rather than protecting data. Thus the traditional de-identification techniques are not applicable in the era of Big Data Since the de-identification technique widespread uses. The tasks of ensuring Big Data security and privacy become more difficult as information is increased. Computer scientists have repeatedly shown that even anonym zed data can often be re-identified and attributed to specific individuals.

7.1 Different Types of Encryption Technique

Table 2. Encryption Technique

Encryption scheme	**Features**
Identity based Encryption	Access control is based on the identity of a user. Complete access over all resources.
Attribute Based Encryption	Access control is based on user attribute. More secure and flexible as granular access control.
Homomorphic Encryption	Computations are performed on the encrypted data and it is very secure.

7.1.1 Identity Based Encryption

An Identity Base Encryption (IBE) is a public key cryptosystem where any string is a valid public key (Savant, 2015). In that case, email address and dates can be a public key. The original motivation for identity-based encryption is to help the deployment of a public key infrastructure. Generally, IBE can simplify systems that manage a large number of public keys from usernames, or simply use the integers {1,....,n} as distinct public keys. Table 2 shows the encryption techniques.

7.1.1.1 Attribute Based Encryption

Attribute based encryption is a type of public key encryption in which the secret key of a user and the ciphertext are dependent upon attributes (ciphertext-policy ABE-CP-ABE).In ciphertext policy attribute based encryption (CP-ABE) a user's private key is associated with a set of attributes and a ciphertext specifies an access policy over a defined universe of attributes within the system (Savant, 2015). A user will be able to decrypt a ciphertext, if and only if his attributes stratify the policy of the respective ciphertext. Policies may be defined over attributes using conjunctions, disjunctions and (k, n) threshold gates, ie, k out of n attributes have to be present. Let us assume that the universe of attributes is defined to be { A, B, C, D} and user1 receives a key to attributes{A,B} and user2 to attribute{D}. If a ciphertext is encrypted with respect to the policy (A□C)□D, then user2 will be able to decrypt , while user1 will not be able to decrypt. CP-ABE thus allows to realize implicit authorization, and it is included into the encrypted data and only people who satisfy the associated policy can decrypt data.

7.1.2 Homomorphic Encryption

Homomorphic encryption is an encryption algorithm which allows specific types of computations to be carried out on plain texts and generate an encrypted result which, when decrypted, matches the result of operations performed on the plain texts. Homomorphic encryption allows complex mathematical operations to be performed on encrypted data without compromising the encryption.

8. CONCLUSION

Big data needs extra requirements for security and privacy in data gathering, storing, analyzing, and transferring. In this paper, we examined studies on big data, classification, big data analytics, cloud based data, technologies used in big data, security and privacy, and different encryption techniques. Big data privacy, safety and security are the biggest issues to be discussed more in future , so new techniques, technologies and solutions need to developed in terms of human computer interactions or existing technologies should be improved for accurate results. It is hoped that this study would help understand the Big Data and its ecosystems better and develop better systems, tools, structures and solutions.

REFERENCES

"Big data tutorial: Everything you need to know". (2015). Retrieved from Searchstorage.techtarget.com.Web. Borovick, Lucinda, Villars, & Richard, L. (2012, April). "The critical role of the network in big data applications .*Cisco White Paper*

Divakar, M., Shrikant, K., & Shweta, J. (2015). Big data architecture and patterns. Retrieved August 1, 2015, from http://www.ibm.com/developerworks/library/bd-archpatterns1/

Lindell, Y., & Pinkas, B. (n.d.). Privacy preserving data mining" in Advances in Cryptology. 36_54.

Marchal, S., Xiuyan, J., State, R., & Engel, T. (2014). A Big Data Architecture for Large Scale security Monitoring. 56-63.

Matturdi, B., & Zhou, X. (2014). Big Data security and privacy A:review, *Big Data, Cloud & Mobile Computing* , 11 (14), 135-145.

Mehta, B. B. (2016). Towards Privacy Preserving Big Data Analytics . International conference on Advanced Computing & Communication Technologies (ACCT - 2016).

Miloslavskaya, N., Senatorov, M., Tolstoy, A., & Zapechnikov, S. (2014). Information Security Maintenance Issues for Big Security-Related Data. Future Internet of Things and cloud (Ficloud), 361-366.

Paul, P. S. (2013). Big data analysis: Challenge and solution . *International conference on cloud , big data and trust* Ranchi ,India: Birla institute of technology.

Saraladevi, B., Pazhaniraja, N., Paul, P. V., Basha, M. S., & Dhavachelvan, P. (2015). Big Data and Hadoop-A study in Security perspective. 50, 596-601.

Savant, V. G. (2015). Approaches to Solve Big Data Security Issues and Comparative Study of Cryptographic Algorithms for Data Encryprtion. *International Journal of Engineering Research and General Science* , Volume 3 (Issue 3).

Tor, K., & Mona, M. (2012). Applying big-data technologies to network architecture . Ericsson Review.

Toshniwal, Raghav, Dastidar, Ghosh, k., Nath, & Ashok. (2015). Big Data Security issue and challenge. International Journal of Innovative in Advanced Engineering (IJIRAE) , volume 2.

Vijey, T., & Aiiad, A. (2015). Big Data Security Issues Based on Quantum Cryptography and Privacy with Authentication for Mobile Data Center. 50, 149-156.

Xuyun, Z., Wanchun, D., Jian, P., Nepal, S., Chi, Y., Chang, L., et al. (2013). Combinig Top-Down and Bottomup:Scalable sub-tree Anonymization over Big Data Using MapReduce on Cloud. 501-508.

Acknowledgement

Geethanjali M was born in Dharmapuri, TN, India in 1983. She received the Master of Science in Computer Science degree from the Periyar University, Salem, TN, India in 2006 and Master of Philosophy in Computer Science degree from the Periyar University, Salem, TN, India in 2007. She is currently pursuing the Ph.D Degree with the Department of Computer Science, Periyar University, Salem. Her research interests include data mining, big data and cloud computing.

Dr Madhubala P was born in Dharmapuri, TN, India. She received Ph.D Degree in Mother Teresa University, Kodaikanal , TN, India. She has 13 years of teaching experience and guiding M.Phil and Ph.D scholars. Her area interests include cloud computing and network security.