Survey on big data security and privacy for Organizations and Government Agencies

Lalita Devi

Research Scholar, Department of Computer Science and Application, Singhania University, Jhunjhunu,

Rajasthan, India

ABSTRACT: The most notably aspect required by the Big Data is Security and Privacy. Different organizations and Government agencies need to share the different kinds of data everyday with high efficiency and effectiveness. So now a day's big data is one of the measure problems to solve, over there researchers are focussing. Survey of big data security and the mechanisms that used to protect and secure also have privacy for big data with an available clouds.
This paper carefully highlights various security issues Big Data analytic faces so far and encourages, further collaborative research for mitigating both security and privacy challenges relating to big data.

KEYWORDS: Big Data, Organizations and Government Agencies with Big Data, Data Providers, Security and Protection, Privacy.

I. INTRODUCTION

Big data refers to datasets that are traditionally not big, but also the high in variety, volume and velocity, which makes them difficult to handle using traditional tools and techniques .But decision makers need only to gain the valuable insight from these huge collection of data sets. Web based services and solutions provide the fast development, processing and to improve various organizations and government agencies. The performance of a manufacturing system must be evaluated and monitored closely to identify variation and disruption in many areas of manufacturing system to facilitate necessary improvement. In 2012, limits on the size of data that could be process in a reasonable amount of time were on the Exabyte of data.
Various Government agencies like Ministry of Statistics, Ministry of Labour Bureau, Ministry of Health and Family Welfare, Ministry of Environment, Forest and Ministry of Education and etc. Each and every day these agencies collect the huge amount of data and store in distributed environment and process. So they data analytics can be used for system performance improvement of system design decomposition cost justify resource allocation decision for improvement.
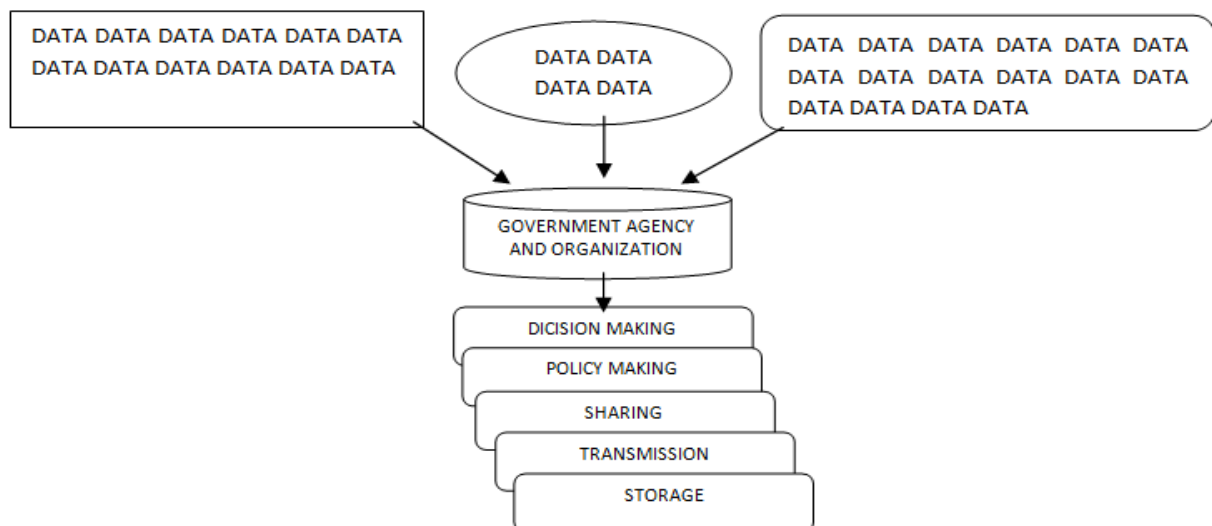
With proper use of big data technologies and application, Government agencies will be able to exploit these data and transformed it into valuable information as well as sharing of information to other agencies. The privacy and security providing such forum for researchers, and developers to exchange the latest experience, research ideas and development on fundamental issues and applications about security and privacy issues in cloud and big data environments. The cloud helps organizations and enables rapid on demand provisioning of server resources such as CPUs, manage, storage, bandwidth, and share and analyze their Big Data in a reasonable and simple to use. The cloud Infrastructure as a service platform, supported by on demand analytics solution seller that makes the large size of data analytics very affordable.

II. BIG DATA FOR ORGANIZATIONS AND GOVERNMENT AGENCIES

Web based services upgrade the system in competitive environment. An absolute design applied to manufacturing system and, is a design methodology to best reflect, understand and control the inherent complexity of large scale integrated system. The desired objective of any system is stability, eventual cost and time reduction and sharing of information is an important approach for Government Agencies to enlarge their views and create values from multiple perspectives. Each and every day these agencies collect the huge amount of data and store in distributed environment and process. Between 2013 and 2020 it will go to 44 trillion GB from 4.4 trillion GB. Moreover the huge amount of data recorded mostly in nonstandard forms which cannot be analyzed using traditional data models and methods. Big Data today have a wide range of challenges but the opportunities are also exists the right decision making, marketing strategies and improved customer relations, better public services and so on. Big data has opened the growing interest to new tools production Beginning with the introduction of Apache Hadoop and Map Reduce and also many open source have been implemented and developed by companies IBM, Oracle, Cloudera, SAP, Teradata, SAS Amazon and many others.
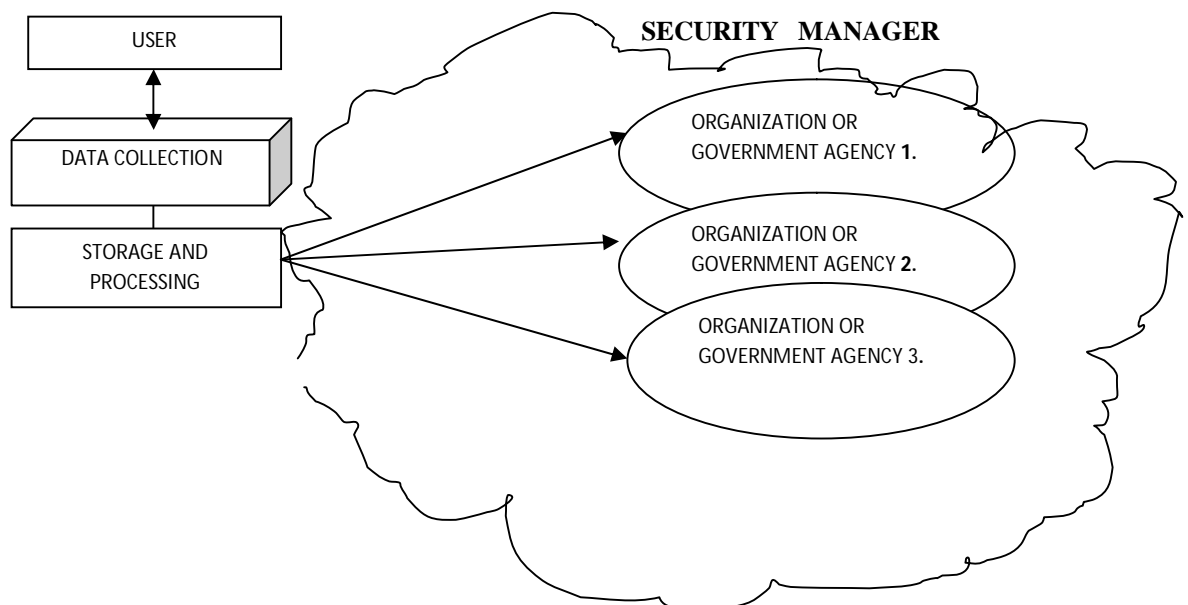

Fig. 1 Illustrate the Architecture of Distributed Data Sharing with Security without Affecting Sensitive Information.

With proper use of big data technologies and applications, organizations will be able to exploit these data and transformed it into valuable information.

2.1 Big basic challenges of unstructured big data –

2.1.1 Storage and processing issues: - Big data analytics fall into two categories: capacity and performance.
These days enormous amount of data have become available on hand to decision makers in an Organizations and in Government Agencies. Organization would thus lose the ability to extract the valuable information from and knowledge, perform detailed analysis, as well as provide new opportunities and advantage to improve their system.
The data transfer from storage need to maintain data integrity before processing for a system.

2.1.2 Privacy and Security: -

- It is the most important issue with big data which is sensitive and includes conceptual, technical as well as legal significance.
- The big data is stored using distributed architecture. So the data is portioned horizontally, vertically, replicated and distributed among multiple nodes. So the data needs to be processed securely.
- The unstructured data on social sites are changing continuously. So it is necessary to capture the changing data for processing.
- The queries are complex and are needs to handle in parallel.
- Due to large size of data, instead of moving data between the different nodes, it is feasible to move the data codes. So the data security is essential.
- Basically we need to store the large amount of unstructured data. So it is necessary to manage the large volumes of data.
- Big data captures the data from various logs, social, media, etc. So we needs to identify who has the right access to the data at what time and from which location.

2.1.3 Data access and sharing of information between different nodes: - if data is used by organization for decision making or manufacturing design on time .so it necessary that it should available on time with consistency. Sharing of data among various organizations must be secured.

2.1.4 Big data analytics: - Today's big data sizes are constantly increasing, currently ranging from a few dozen Terabytes (TB) to many Petabytes (PB) of data in a single data sets. Data analytics is the process of applying algorithms in order to analyze sets of data and extract useful information and unknown hidden patterns, relationships, and information. Big data analytic is where advance analytics techniques are applied on big data sets.

 Naturally the organizations and government agencies benefits are derived from analyzing the more complex and huge data sets that require real time and near real time capabilities. The storage of big data chances to increase malicious attacks and abuse activities. The attacks can take place during the exchange of information and transmission of data. The unexpected operations can be done on remote site.

Our proposed mechanism aims to encrypt all data and sharing the data among the various organizations and Government agencies without causing big overhead or latency.fig. 1 Illustrate the architecture of distributed data sharing with security without affecting sensitive information.

2.1.5 Heterogeneity and Incompleteness: - Pointing out the acquisition of big data from various sources with variety of structures, structuring these data is almost impossible before data analysis. Computer systems are known to work more efficiently if they can store these data all identical in size and structure and complete.

III. BIG DATA PRIVACY AND SECURITY

Basically term big data refers to the three V's in which privacy and security issues are magnified.

- Volume: Huge collection or amount of data
- Velocity: speed of coming data from various ways

- Variety: Different kind of data eg. Audio, video, textual data.

Big Data could significantly improve government policymaking, social-welfare programs, and scholarship.

Even Google's search engine is not immune. Despite being driven by an enormous amount of data overseen by some of the world's top data scientists, its results are susceptible to "search-engine optimization" and manipulation, such as "Google bombing," "spamdexing," and other methods serving parochial interests. The sharing, linking and analysis of data across government to provide new insights for the purposes of supporting:-

- Policy development
- System planning
- Resource allocation
- Performance monitoring

Fundamental tension between big data and some basic principles of privacy personal information (PI) should be collected directly from the individual.

Personal Information (PI) should only be used for the purpose for which it was collected. Big data involves information that has been collected indirectly used for a purpose which may not have been intended at the time of collection. Generation of new PI not collected directly from the individual. Big Data implementation as needed with the result that the entire data storage layer needs security protection.
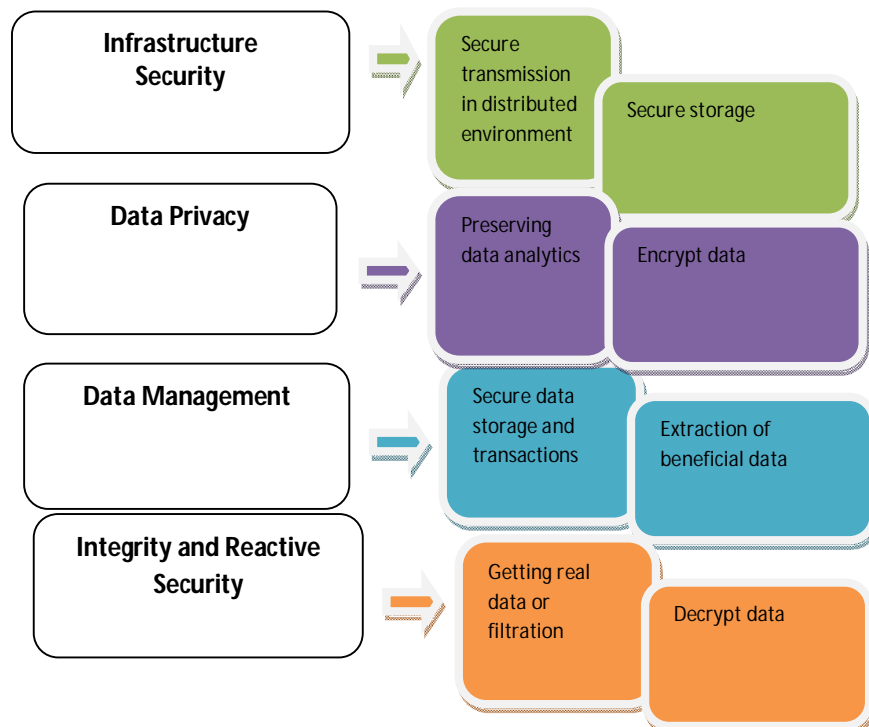
There are many types of protection and security used such as above in figure.

3.1 Encryption and Decryption: To centrally manage encryption and key management technique that enables compliance and is transparent to the processes, applications and to the end users. No one can access the confidential information until the key is not known.

3.2 Hadoop: Hadoop is an open source framework used for big data in distributed environment. Similar to data residing in a local file system of personal computer system, in Hadoop, data resides in a distributed file system which is

called as a Hadoop Distributed File system. Processing model is based on 'Data Locality' concept wherein computational logic is sent to cluster nodes (server) containing data. Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers.

Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at low cost.

3.2.1 Features of Hadoop:

- Suitable for Big Data Analysis
- Scalability
- Fault Tolerance

3.3 MapReduce: MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

MapReduce composed of two thing, Map() called procedures that perform filtering and sorting on data sets and Reduce() method that perform a summary operations.

MapReduce usually split the input data sets into independent chunks and all are processed into parallel manner.

IV. **CONCLUSION**

Big Data is changing the way we perceive our world. Companies are using big data analysis to target marketing at very specific demographics. We can make better advances in many scientific disciplines and profitability for many enterprises. To manage and to process the big data, more efforts are required. Security is one of the challenges that arise when systems try to handle the concept of big data. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. There is immense scope in Big Data and a huge scope for research and Development. There are so many Technologies, such as the Apache Hadoop distributed computing framework and NoSQL databases have challenge of very large and unwieldy datasets. And now another technology, already at work behind the scenes, could grow in importance in the coming years.

REFERENCES

1. A community White paper developed by leading researchers across united states, "Challenges and Opportunity with Big Data", Feb. 2012.
2. A Cloud Security Alliance Collaborative research, "Expanded Top Ten Big Data Security and Privacy challenges" , April 2013.
3. http://www.businessweek.com/articles/2013-08-07/the-futureof-big-data-apps-and-corporate-knowledge.
4. http://www.qubole.com/big-data-cloud-database-computing/
5. https://cloudsecurityalliance.org/media/news/csa-releasesthe-expanded-top-ten-big-data-security-privacy-challenges/
6. http://data-informed.com/cloud-computing-experts-detailbig-data-security-and-privacy-risks/
7. http://www.vormetric.com/data-securitysolutions/applications/big-data-security
8. J. Feng, Y. Chen, W.-S. Ku, and P. Liu, "Analysis of Integrity Vulnerabilities and a Non-repudiation Protocol for Cloud Data Storage Platforms," the 2nd International Workshop on Security in Cloud Computing (SCC 2010), in conjunction with ICPP 2010, San Diego, California, USA, Sept. 14, 2010.
9. T. Vijey, A. Aiiad, "Big Data Security Issues Based on Quantum Cryptography and Privacy with Authentication for Mobile Data Center", Procedia Computer Science, vol. 50, pp. 149–156, 2015.
10. B. Matturdi, X. Zhou, S. Li, F. Lin, "Big Data security and privacy: A review", Big Data, Cloud & Mobile Computing, China Communications vol.11, issue: 14, pp. 135 – 145, 2014.
11. C.L.P. Chen, C.Y. Zhang, "Data Intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences , vol. 275, pp.314-347, 2014.
12. N. Miloslavskaya, M. Senatorov, A. Tolstoy, S. Zapechnikov, "Information Security Maintenance Issues for Big Security-Related Data", Future Internet of Things and Cloud (FiCloud), pp. 361 – 366, Barcelona, 2014.
13. R. Bryant, R. H. Katz, and E. D. Lazowska, "Big-data computing: creating revolutionary breakthroughs in commerce, science and society," December 2008.
14. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, and M. G. Institute, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute San Francisco, 2011.