Volume: 08, Issue: 05 Pages: 5-9 (2017) Special Issue

Security Perspectives on Deployment of Big Data using Cloud: A Survey

R.Kalaivani
Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi
Email:kalaisudha.2005@gmail.com

ABSTRACT

As everything was digitized today, tons of petabyte and exabyte of data were generated by every organization, industry, business function and individual throughout the world. Because of this scenario, conventional data was slowly turned into Big Data. As user could not afford all infrastructures to support Big Data technology, Cloud Computing technology was boomed as user dream come true by performing massive-scale and complex computing. The major challenging issue in deploying Big Data using Cloud was security. All our expensive data were present at various vendor clouds; there might be huge chances for our data to be exploited knowingly or unknowingly by others. So through this paper, I concentrated on providing security to big data which was stored on cloud.

Keywords: **Big Data, Cloud, Security**

I. INTRODUCTION:

As air occupies everywhere, now digital appliances present everywhere. Usage of digital appliances increases in order of exponential throughout the year. Nowadays, Social media such as Facebook, Whatsapp, Twitter, etc were started used by everybody which produces millions and trillions of data every day. Big data is eliciting attention from the academia, government, and industry. Big data are characterized by three aspects: (a) data are numerous, (b) data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed rapidly. Moreover, big data is transforming healthcare, science, engineering, finance, business, and eventually, the society [1]. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opened a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources. This paradigm is being popularly termed as Big Data [2]. Despite of its popularity, deploying Big Data Technology was tedious as well as time consuming process. Various researches were done to optimize the storage and retrieval of data in useful pattern which was used to identify users' requirements and demands. But the persistent increase in data gave a big confront to the researchers.

Cloud computing has been revolutionizing the IT industry by adding flexibility to the way IT is consumed, enabling organizations to pay only for the resources and services they use. In an effort to reduce IT capital and operational expenditures, organizations of all sizes are using Clouds to provide the resources required to run their applications. Clouds vary significantly in their specific technologies and implementation, but often provide infrastructure, platform, and software resources as services [3][4]. Cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools [5].

In this paper, Section 2 presents Big Data definition; characteristics and classification of Big Data; Section 3 covers Cloud Computing; the relation between big data and cloud was discussed in Section 4; Section 5 deals the challenges in security; Last section gave the conclusion of the study.

II. BIG DATA

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of ―Big Data and the Next Wave of Infra Stress. Many Researchers and organizations have tried to define Big Data in different ways. Gartner defines Big Data are high -volume, high-velocity and high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization [6] [7].Several definitions of big data currently exist. For instance, [8] defined big data as "the amount of data just beyond technology's capability to store, manage, and process efficiently". Meanwhile, [9] and [10] defined big data as characterized by three Vs: volume, variety, and velocity. IDC also defined big data technologies as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis."

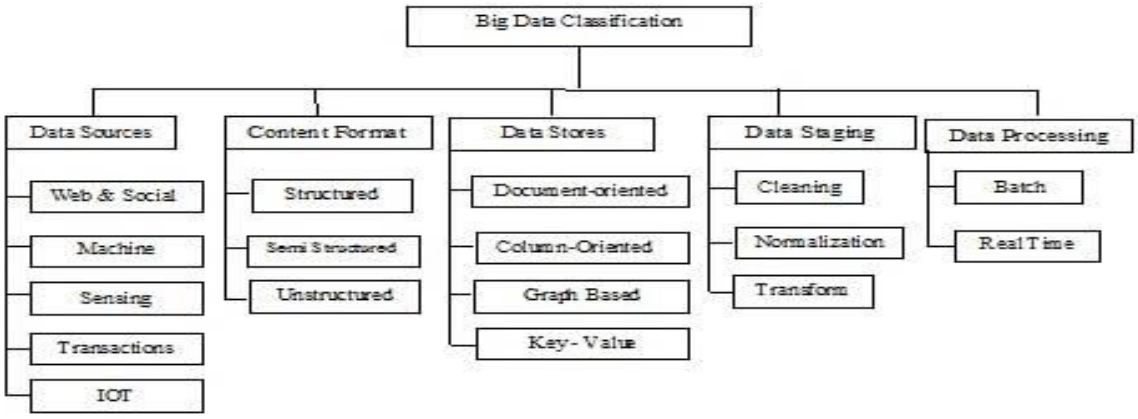

Fig 1: Big data classification

Stage/Year		Characteristics	Examples
Batch Processing	2003-2008	Big amount of data is collected, entered, processed and then the batch results are produced. Distributed file systems (DFS) are used for fault-tolerant and scalability. Parallel programming models such as MR are used for efficient processing of data.	GFS, MR, HDFS, Apache Hadoop
Ad-hoc(NoSQL)	2005 - 2010	Support random read/write access to overcome shortcomings of DFS that are appropriate for sequential data access. NoSQL databases solve this issue by offering column based or key-value stores, in addition to support for storage of large unstructured datasets such as documents or graphs.	CoachDB, Redis,Amazon DynamoDB, Google Big Table, HBase, Cassandra, MongoDB
SQL-like	2008 - 2010	Simple programming interfaces to query and access the datastores. This approach provides functionalities similar to the traditional data warehousing mechanisms.	Apache Hive/Pig, PrestoDB, HStore, Google Planner
Stream Processing	2010 2013	Data are pushed continuously as streams to servers for processing before storing them. Streaming data usually have unpredictable incoming patterns. Such data streams are processed using fast, fault-tolerant, and high availability solutions.	Hadoop Streaming, Google Big Query, Google Dremel, Apache Drill, Samza Apache Flume/Hbase, Apache Kafka/ Storm
Real-time Analytical Processing	2010-2015	Automated decision making for streams that are generated from the machine-to-machine applications or other live channels. This architecture helps to apply real -times rules for the incoming events and existing events within a domain.	Apache Spark, Amazon Kinesis, Google Dataflow

Table 1: Evolution of Data throughout time

Table 1 summarizes the big data technologies from batch processing in 2000 to present with most significant stages and products [11].

Big Data are classified into different categories to better understand their characteristics. It was depicted here as Fig 1. Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in various formats. Most popular is the relational database that comes in a large number of varieties [12]. As the result of the wide variety of data sources, the captured data differ in size with respect to redundancy, consistency and noise, etc.

III. CLOUD COMPUTING

Cloud computing is becoming a reality for many businesses, with private cloud deployments often leading the way. Cloud technology is maturing and addressing barriers to adoption with improvements in security and data integration, while IT organizations are evolving to support cloud services delivery. As a result, businesses are demonstrating growing trust in cloud delivery models. For example, a 2013 survey from Ubuntu found that 55 percent consider the cloud ready for mission-critical workloads [13].

Figure 2: Cloud Services

Figure 2 describes the services provided by the cloud. Organizations continue to store more and more data in cloud environments, which represent an immense, valuable source of information to mine. Plus, clouds offer business users scalable resources on demand. Hadoop software provides the high-performance computing power needed to analyze vast amounts of data efficiently and cost-effectively. Running Hadoop in virtualized environments continues to evolve and mature with initiatives like VMware's open-source project Serengeti*, among others.

IV. How Cloud and Big Data Related:

Cloud computing and big data are now become as twin technology. Big data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner. Cloud computing provides the underlying engine through the use of Hadoop, a class of distributed data-processing platforms. The use of cloud computing in big data is shown in Fig. 3 .Large data sources from the cloud and Web are stored in a distributed fault-tolerant database and processed through a programming model for large datasets with a parallel distributed algorithm in a cluster. The main purpose of data visualization, as shown in Fig. 3, is to view analytical results presented visually through different graphs for decision making. Big data utilizes distributed storage technology based on cloud computing rather than local storage attached to a computer or electronic device. Big data evaluation is driven by fast-growing cloud-based applications developed using virtualized technologies. Therefore, cloud computing not only provides facilities for the computation and processing of big data but also serves as a service model [14]. Cloud computing infrastructure can serve as an effective platform to address the data storage required to perform big data analysis. Cloud computing is correlated with a new pattern for the provision of computing infrastructure and big data processing method for all types of resources available in the cloud through data analysis. Several cloud-based technologies had to cope with this new environment because dealing with big data for concurrent processing had become increasingly complicated [15].

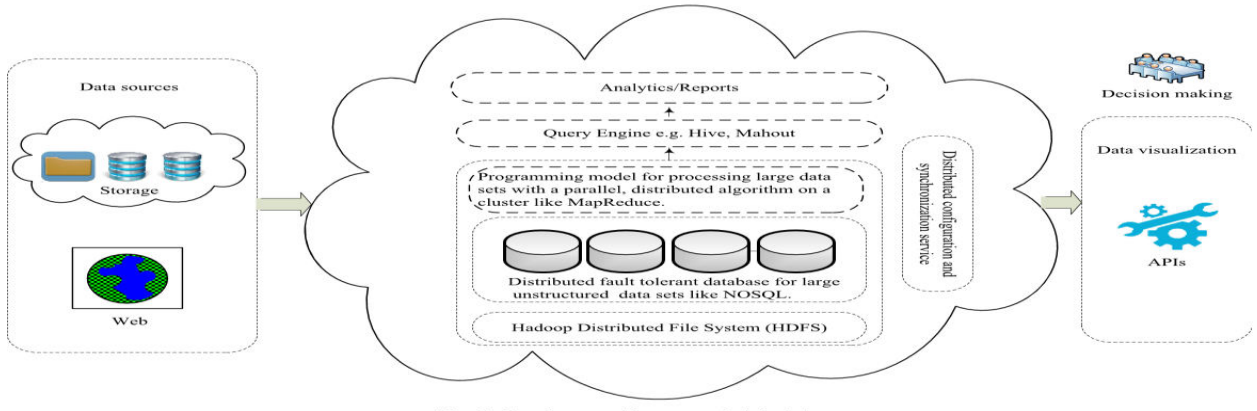

Fig 3: Relationship between Cloud and Big Data

V. Security Challenges in Cloud and Big Data

The security issues associated with cloud computing devices and environments can be categorized into the following: network level, user authentication level, data level, and generic issues as depicted by [16] [17].

.•Network level: The challenges associated with network level will include issues with network protocols and network security, such as distributed nodes, distributed data, Internode communication.

•User Authentication level: The issues and challenges associated with user authentication level includes encryption/decryption techniques, authentication methods which may include issues with administrative rights for nodes, authentication of applications and nodes, logging etc.

•Data level: The issues and challenges associated with data level will include data integrity and availability issues such as data protection and the distribution of data.

•Generic types: The issues and challenges associated with general level security issues include issues with traditional security tools, and use of different technologies.

Since Cloud Computing Technologies was combination of various technologies, I listed few of the security measures that would protect Big Data in Cloud environment.

a. File Encryption:

Since the data is present in the machines in a cluster, a hacker can steal all the critical information. Therefore, all the data stored should be encrypted. Different encryption keys should be used on different machines and the key information should be stored centrally behind strong firewalls. This way, even if a hacker is able to get the data, he cannot extract meaningful information from it and misuse it. User data will be stored securely in an encrypted manner.

b. Network Encryption

All the network communication should be encrypted as per industry standards. The RPC procedure calls which take place should happen over SSL so that even if a hacker can tap into network communication packets, he cannot extract useful information or manipulate packets.

c. Logging

All the map reduce jobs which modify the data should be logged. Also, the information of users, which are responsible for those jobs should be logged. These logs should be audited regularly to find if any, malicious operations are performed or any malicious user is manipulating the data in the nodes.

d. Software Format and Node Maintenance

Nodes which run the software should be formatted regularly to eliminate any virus present. All the application software and Hadoop software should be updated to make the system more secure.

e. Nodes Authentication

Whenever a node joins a cluster, it should be authenticated. In case of a malicious node, it should not be allowed to join the cluster. Authentication techniques like Kerberos can be used to validate the authorized nodes from malicious ones.

f. Rigorous System Testing of Map Reduce Jobs

After a developer writes a map reduce job, it should be thoroughly tested in a distributed environment instead of a single machine to ensure the robustness and stability of the job.

g. Honeypot Nodes

Honey pot nodes should be present in the cluster, which appear like a regular node but is a trap. These honeypots trap the hackers and necessary actions would be taken to eliminate hackers [18].

VI. Conclusion

The usage of mobile phones, social media, industries, organization etc leads to the constant increase in size of data. As data size increases, need for infrastructure, technologies had also hiked. Since data became more precious than any other resources, we are in great need of providing security to safeguard our data from security breaches and vulnerabilities. Even though various methods were already available, we had to still concentrate more on providing security to data resided in cloud.

References:

1. http://dx.doi.org/10.1016/j.is.2014.07.006

2. http://dx.doi.org/10.1016/j.jpdc.2014.08.003

3. R.Buyya, C.S.Yeo, S.Venugopal, J.Broberg, I.Brandic, Cloud computing and emerging ITplatforms:Vision, hype,and reality for delivering computing as the 5th utility, *Future Gener. Comput.Syst.* 25(6)(2009)599–616

4. M.Armbrust, A.Fox, R.Griffith, A.D.Joseph, R.H.Katz, A.Konwinski, G.Lee, D.A.Patterson, A.Rabkin, I.Stoica, M.Zaharia, *Above the Clouds: A Berkeley View of Cloud Computing, Technical report UCB/EECS-2009-28,* Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, USA (February2009).

5. Swati Batra, Dr. A.K Sharma,"A Survey on Security of Big Data on Cloud",*International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4,* Issue 5, May 2016

6. Big Data: science in the petabyte era, *Nature 455* (7209):1, 2008

7. Reena Singh, Kunver Arif Ali, "Challenges and Security Issues in Big Data Analysis", *International*

Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 1, January2016, ISSN(Online): 2319-8753,ISSN (Print):2347-6710

8.J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H.Byers, *Big data: The next frontier for innovation, competition, and productivity*, (2011)

9.P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan,*Harness the Power of Big Data The IBM Big Data Platform* [McGraw Hill Professional, 2012]

10. J.J. Berman, *Introduction, in: Principles of Big Data*, Morgan Kaufmann, Boston, 2013, xix–xxvi (pp)

11. S. Rusitschka and A. Ramirez, "*Big Data Technologies and Infrastructures*." http://byte-project.eu/research/, Deliverable D1.4, Version 1.1, Sept. 2014.

12.J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, *Big data for dummies*,For Dummies (2013)

13.Ubuntu 2013 Server and Cloud Survey. Ubuntu Server (September 10, 2013)

14.I.A.T. Hashem et al. / *Information Systems* 47 (2015) 98–115,The rise of "big data "on cloud computing: Review and open research issues

15.C. Ji, Y. Li, W. Qiu, U. Awada, K. Li, Big data processing in cloud computing environments, Pervasive Systems, Algorithms and NetWorks (ISPAN), 2012,in: Proceedings of the *12th International Symposium on, IEEE*, 2012, pp. 17–23.

16.Venkata N. I., Sailaja A., and Srinivasa R. R. (2014) Security Issues Associated With Big Data in Cloud Computing,*International Journal of Network Security & Its Applications (IJNSA), Vol.6*, No.3, May 2014 DOI:10.5121/ijnsa.2014.6304 45

17.Saranya A, MuthuKumar, V.P. (2015) Security issues associated with big data in cloud computing *International Journal of Multidisciplinary Research and Development Volume: 2,* Issue: 4,580-585 April 2015www.allsubjectjournal.come-ISSN: 2349-4182 p-ISSN:2349-5979

18.Santosh Kumar Satapathy ,Santosh Kumar Moharana, Avay Kumar Ojha, "Implication of Security Issues Associated With Big Data In Cloud Computing " , *International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02*, Issue 04; April-2016[ISSN: 2455-1457]