Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 2818-2823

Research Article

A prodigal paradigm for the solution of issues and challenges which leads in Big data security

Himani Sivaraman[1], M.Manchanda[2],Sanjay Jasola[3], Kamlesh Purohit[4], Amit Gupta[5]

hsivaraman@gehu.ac.in[1], mmanchanda@gehu.ac.in[2], sjasola@yahoomail.com[3], kamleshchandrapurohit@geu.ac.in[4], agupta@gehu.ac.in[5]

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: Recent technological development such as social network, cloud computing, and data analytics make possible to collect large amount of data. Data security and privacy are critical aspects of concern. However , there are issues in securing and protecting this data.Big data is like a two edged sword, as it bring convenience of handling the huge amount of data along with it creates certain risks for the analysts. It is believed in the process of data collection , storage, manipulation, presentation, the data leak- age can exists as unavoidable ghost. The big data architecture being distributive in nature, undergo portioning , replication and distribution of this data on thousands ofdata processing nodes for the distribution of this .How to make the data security andprivacy more significantly strong and to stop the data leakage in the stages of the dif-ferent algorithm has become a essential goal and most important research challenge for researchers and academicians. This paper discusses and review the issues and thekey factors what is to be taken in account while undergoing development of securedbig data solutions. It can increase the performance of impact and resultant solutions for these crisis. The traditional methods of privacy cannot be that much successful inproviding solution to this arising situation.

Keywords: Big data, analytics, Hadoop, MapReduce, Data-privacy, data-security, architec-ture, risk , big data encryption. Secure computation

1. Introduction

Technological advancement in this area of research, the novel and innovative applicationsuch as sensors, mobile devices, cloud computing systems, social media , internet of things etc, make it possible in the role of data collection, storage and processing. The Big Data is being distributive in nature, it contributes in the distribution, collection, replication and partitioning of this humongous data. This distributive nature leads to the support of many Big Data analytical programming frameworks like continuous streaming of data in paral-lel and real time environment(Antignac and Le Métayer (2014)) Many recent organizations are engaged in the development of technologies which have the control , of this replication and distribution of data in different nodes in the real time environment. The data sourcesare not confined to the traditional methods and ways so these companies are using tech-niques and tools for analyzing and storage of this zettabyte of data which is including

2. Research paper for review

The social media feed , live streaming and web logs etc. It leads to the complexity and criticality of classifying the information. The security and risk factors of this huge data are not today's concern but this have always been a favorite topic of research for many organizations. The main motive of this is how to manage the computation of this data and managing it at the same time. The security of data into this broad spectrum con- fined to areas like data confidentiality, availability and its integrity maintenance. The Data confidentiality refers to the scenario of unauthorized access towards this data, whereas the availability is for the access of the correct and authorized users. The data integrity isbasically the data trustworthiness towards being error free(Bahri, Carminati, and Ferrari (2015)). The big data categorizes data by four V's; velocity, variety, veracity, volume. These characteristics make a unique feature for the recognition of big data; these characteristics lead to the privacy techniques and security concerns(Bahri, Carminati, and Ferrari (2016)). These characteristics also leads to the key research challenges related to data protectionand sanctity of it. In this article we are considering over some challenges and suggestingcertain remedial methods in terms of tools and techniques. The main challenges in termsof big data which is widely known are:

- Insecure computation power
- Filtration of Input and Validation
- Granular Access Control

- Insecure Data Storage
- Privacy concern

These challenges in the present scenario can be considered and can be taken into accountby many big data tools and techniques

3. Research review

Several techniques were acquired for data confidentiality and privacy maintenance likecryptographic techniques data structures(Batini, Scannapieco, et al. (2016)) that hide the data access pattern. One techniques which is suggested is making difficult in data link-ing; specific data records to specific individual. Many recommendations were made intaking account of the data processing and computation strategies Bertino (2012)). Some semi automated policy integration (Bertino and Ferrari (2018)) were also suggested by the researchers. An initial pioneering approach was proposed(Bahri et al. (2015))that asso- ciates with each data item a set of possible purposes, from an ontology of purposes, for which the data can be used. Many researchers suggested data anonymization(Bertino, Ja- hanshahi, Singla, and Wu (2021)), and some of them suggested to have advance accessdata node models (Bhandari, Hans, and Ahuja (2016)). In recent times many suggestedthe IOT based data driven crypto methods(Bertino (2012), which are mainly in-terms of the data generated by the internet of things devices. Today one of the most used defi- nition of data privacy is due to Allan Westin that defined data privacy as the " claim of individuals, groups or institutions to determine for themselves when , how and to what extent information about them is communicated to others".(Ceravolo et al. (2018)) . Veryoften we provide a platform for data privacy as the same requirement for data confiden-tiality but more reference to personal data. In(Colombo and Ferrari (2015)), the descriptionof the data security has been considered as a method for providing data on to the cloud computing as well.

4. Big data architecture

The Big data Architecture comprises of many skills like; development of reliable and auto- mated data pipeline. Actually, if we can precisely say that there is no particular standard-ized architecture available for big data; being it a new field of research. Characteristics like latency, velocity, volume, veracity,scalability , fault tolerance just become key and important features that make it mandatory for choosing big data architecture. Many other intrinsic attributes like autotiering, easy shift can also undergo security concerns. The Fig- ure1, describes the distributiveness of big data architecture. The recent architecture pro- visions real time computation capability. The data sources can be done with many other online sources apart from traditional ways(Essakimuthu, Ganesh, Krishnan, and Robinson (2021)). The social media and other web logs also contributes to these collection of input data sources. Map-Reduce has given this framework a more flexible and more powerful execution programming paradigm. The program divides into many data nodes execute the respective data-node and finally reduce it to a single set result.

5. PARALLEL AND POWERFUL PROGRAMMING PARADIGM- mapreduce

MapReduce framework has the ability to process your data with distributed computing. It is a programming model and a distributed computing framework to process extremelylarge data. It work on to write auto scalable distributed applications in a cloud environ- ment. In the MapReduce model as shown in Figure 2, programmers have to reduce an algorithm into iterations of MAPPER and REDUCER functions. Writing an algorithm onlyconsisting of these two functions can be a complicated task, but MapReduce framework is able to automatically scale and parallelize such algorithms. The framework take caresof partitioning the input data, scheduling, synchronizing and handling failures, allowing the programmers to focus more on developing the algorithms and less on the backgroundtasks. The MapReduce is the most sustainable and powerful programming paradigm in case of bigdata.Take in consideration that we have 16 TB of data coming in continuation.The data is divided in 128MB Chunks(Standard Hadoop method of block creation) Thenthe program will divide them into 82000 Maps (will create the mapper class for this much)

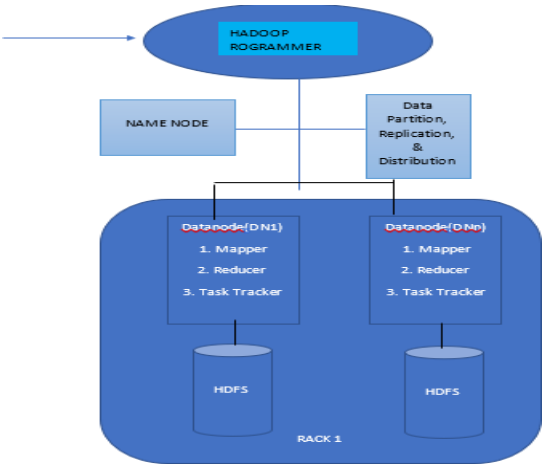

Figure 1: Big Data Architecture

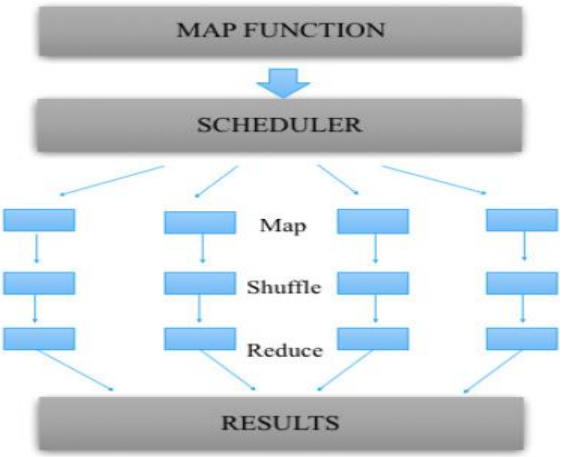

Figure 2: MapR-Architecture

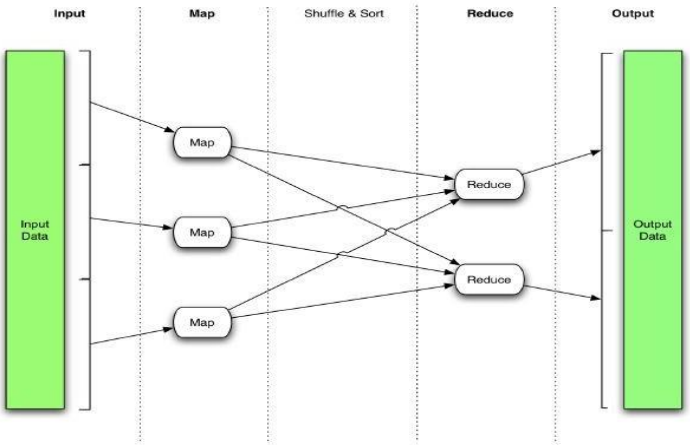

Figure 3: MapReduce

and have a class concurrently running on data processing nodes. Map Reduce frameworkas depicted in Fig 3 is required by Hadoop where the program get divided into multiple map and then get execute it into multiple data nodes and shuffle them to make a merger to- gether into a single set. Many other topologies exist for data processing and managementin the field of big data computation.

4.1 CHALLENGES IN TERMS OF BIG DATA SECURITY

The security mechanism in the big data technologies is not generally weak. The Figure4 , describes the point-to point security mechanism is one of the best in-terms of big datahandling(Husain, Zeebaree, et al. (2021)). Acquiring the robust security mechanism for the purpose by using the features like parallelism ,auto-tiering etc.

- Privacy Concern in Data Mining: The data mining concepts have many issues con- cerning the privacy and the analytical results also involves many challenges example- disclosure of information, public- private key disclosure
- Insecure data Storage: Authentication and authorization of data is main concern . As the data is stored and managed by the data nodes. The insecure data computation

,authentication, authorization and encryption of data to lesser secure medium .

- Computation Insecurity: As mentioned in Figure 4,the untrusted computational pro- gramming paradigms which is used by the attacker in order to extract or turnout sensitive and confidential information from data resources. It can not only cause ainformation leak and also corrupt the data leading into the in-valedictory results tothe prediction or analysis.
- In DOS input validation and filtering: The Denial of Service(DOS) will also havea effect of disabling the property of using and accessing of massively parallel pro-

gramming languages in input validation. As Big data needs in collection to input from variety of data sources it therefore need a more quite importantly and manda-tory valedictory input. Along with a filtering of malicious data and rouge data fromthe good ones.

These above mentioned challenges can be taken into account and can be rectified by certain solutions like cryptography(Pandit, Deshpande, and Karmarkar (2013)) , secure computa- tional data storage(Das and Dash (2021)), implementation of comprehensive input valida- tion etc. The Big Data processing requires a faster response time for its computation andgetting added up in the security implication(Malhotra, Sethi, and Mittal (2021)). In the below section of this manuscript, we are discussing two of the solutions.

4.2 CRYPTOGRAPHIC SOLUTIONS FOR BIG DATA SECURITY

In Hadoop there is an absence of algorithm to encrypt or decrypt the on board data viz.local as well as the HDFS file system. Hadoop works on the Linux platform so it takesthe Linux local system as the storage which is temporary storage. After the processing of map-reduce task the output of the map-reduce gets it's into the local as well as the HDFS (with the help of user). Hadoop has only one end to end security system which is the Ker- beros. Kerberos is the service which basically keeps track of the user access to the particu- lar service and provides restriction policy(Wang et al. (2014)). Kerberos provides only the security gateway to restrict the access of unwanted or unauthorized users to the Hadoop environment and the services. This is basically a policy manager of the Hadoop. Let's take an assumption, what if an unauthorized user gets the credentials of the Hadoop environment, then the entire Hadoop system gets compromised and it will leads to data theft or data loss(Colombo and Ferrari (2015)) To secure this loop hole we can create a system of RSA+AES encryption and decryption algorithm, so that even when the Hadoop system gets compromised, the data in the HDFS or in local has no impact of this compromised situation(Pandit et al. (2013)) After the encryption and decryption the data gets saved inthe HDFS. And it only provides write once and read many times. But there is a simple condition. The user should have the combination of keys as well as some supported filesto open (decrypt) that file for reading purpose(Zhang (2018)). This is an automated process

i.e. if the user close this file after reading, and want to read the file again, he/she needs to provide the keys and the file combination to read it again.

6. Secure computation and data storage

Segregation of your sensitive and confidential data is important in big data privacy paradigm

. The filtering of almost every internal and external sources have to be mandatory. The proper evaluation of key input validation and filtration features of respective big data sources and solutions is required to oversee

whether it can scale up the data requirements and security issues. There are generally two ways of preventing attacks : securing data when insecure mapper is present and securing the mapper in altogether manner.

Figure 4: Files creation :Encryption and decryption

Enabling data node encryption for sensitive data marks a major successful scenarioalong with the verification of proper configuration of API security of all the components involved(Sarosh, Parah, Bhat, and Muhammad (2021)) in the framework. The Analytical algorithms which are used in analysis and prediction such as designing of data , classifica- tion and regression have to be verified timely so that the sensitive data will be sensitized timely . It will definitely reduce the rate of disclosure of the sensitive and confidentialdata. It is important to establish proper guidelines and recommendations for the preven- tive measures used in big data analytics and it has to be pen tested

7. Conclusion

As big data is trending , so are trending the security issues related to it . There is no application which can be imagined with out creating new forms of data , operating newdata driven algorithms , and consuming a humongous amount of data. With the facilitiesof having new data storage cheaper and cloud environment more capable of sharing and storing systems and analytical applications , it is important to have a solid guidelines and recommendation system for the data security. The real- time monitoring techniques , the heterogenous scenario of data producing , it leads to the ad-hoc approaches for securityand privacy . This manuscript have tried some specific aspects of the vulnerable areas in big data processing infrastructure. The solutions like cryptographic solution for bigdata security along with the secure computation can be one of the suggested solutions for the data security issues . The researches suggested that the solution to theses data security cannot be permanent but can only be modified. In future

References

1. Antignac, T., & Le Métayer, D. (2014). Privacy by design: From technologies to architec- tures. In Annual privacy forum (pp. 1–17).
2. Bahri, L., Carminati, B., & Ferrari, E. (2015). Cards-collaborative audit and report data sharing for a-posteriori access control in dosns. In 2015 ieee conference on collaboration and internet computing (cic) (pp. 36–45).
3. Bahri, L., Carminati, B., & Ferrari, E. (2016). Coip—continuous, operable, impartial, and privacy-aware identity validity estimation for osn profiles. ACM Transactions on the Web (TWEB), 10(4), 1–41.
4. Batini, C., Scannapieco, M., et al. (2016). Data and information quality. Cham, Switzerland:
5. Springer International Publishing. Google Scholar, 43.
6. Bertino, E. (2012). Data protection from insider threats. Synthesis Lectures on Data Manage- ment, 4(4), 1–91.
7. Bertino, E., & Ferrari, E. (2018). Big data security and privacy. In A comprehensive guide through the italian database research over the last 25 years (pp. 425–439). Springer.
8. Bertino, E., Jahanshahi, M., Singla, A., & Wu, R.-T. (2021). Intelligent iot systems for civil infrastructure health monitoring: a research roadmap. Discover Internet of Things,1(1), 1–11.
9. Bhandari, R., Hans, V., & Ahuja, N. J. (2016). Big data security–challenges and recommen- dations. International Journal of Computer Sciences and Engineering, 4(1), 93–98.
10. Ceravolo, P., Azzini, A., Angelini, M., Catarci, T., Cudré-Mauroux, P., Damiani, E., ... others (2018). Big data semantics. Journal on Data Semantics, 7(2), 65–85.
11. Colombo, P., & Ferrari, E. (2015). Privacy aware access control for big data: A research roadmap. Big Data Research, 2(4), 145–154.
12. Das, M., & Dash, R. (2021). Role of cloud computing for big data: A review. Intelligent and

13. Cloud Computing, 171–179.
14. Essakimuthu, A., Ganesh, R. K., Krishnan, R. S., & Robinson, Y. H. (2021). Enhanced hadoop distribution file system for providing solution to big data challenges. Further Advances in Internet of Things in Biomedical and Cyber Physical Systems, 71–83.
15. Husain, B. H., Zeebaree, S. R., et al. (2021). Improvised distributions framework of hadoop: A review. International Journal of Science and Business, 5(2), 31–41.
16. Malhotra, J., Sethi, J. K., & Mittal, M. (2021). Analysis of big data using two mapper files in hadoop. International Journal of Security and Privacy in Pervasive Computing (IJSPPC), 13(1), 69–77.
17. Pandit, A., Deshpande, A., & Karmarkar, P. (2013). Log mining based on hadoop's map and reduce technique. International Journal on Computer Science & Engineering, 270– 274.
18. Sarosh, P., Parah, S. A., Bhat, G. M., & Muhammad, K. (2021). A security management framework for big data in smart healthcare. Big Data Research, 100225.
19. Wang, X. S., Nayak, K., Liu, C., Chan, T. H., Shi, E., Stefanov, E., & Huang, Y. (2014). Oblivious data structures. In Proceedings of the 2014 acm sigsac conference on computer and communications security (pp. 215–226).
20. Zhang, D. (2018). Big data security and privacy protection. In 8th international conference
21. on management and computer science (icmcs 2*018)* (pp. 275–278).