ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

A Novel Approach for Identification of Hadoop Cloud environment

[1] Swathi Agarwal
[1] Assistant Professor, Department of Information Technology,
Anurag Group of Intuitions, Hyderabad, Telangana, India

***Abstract* - Because of the advanced propensities inside the teach ofscience and innovation brought about the trendsof viable data switch, capacity of dealing with colossaldata and the recovery of information effectively. On account that the datathat is spared is expanding voluminously, strategies toretrieve relative comprehension and assurance associatedconcerns are to be tended to viably to quiet thisbulk information. Moreover with rising benchmarks of mammoth information,these security issues are a testing undertaking. This paperaddresses the constraint of agreeable information switch utilizing theprinciples of learning mining in cloud condition making use ofhadoopmapreduce. In light of the experimentation achievedoutcome are examined and spoken to with perceive totime and space unpredictability when thought about hadoopwith non hadoop approach.**

Keywords — Big Data, Hadoop, Mapreduce, CloudComputing, Temporal Patterns

I. INTRODUCTION

The contemporary mechanical attributes saw the capacity of enormous data and approachs exceptional towards productive recoveries. For the reason that this data is available are surmounting, wellbeing breaks and maintaining the protection is a principal inconvenience. These security issues are much more troublesome while in light of the fact that the information moves in cloud environment or parallel preparing structures [1]. To deal with this information viably thoughts of mapreduce [2] is packed in the writing. This is because of its abilities of faulttolarence and versatility in conjunction with straightforwardness.

One other central preferred standpoint of featuring the mapreduce idea is it permits the parallel preparing conditions which help not specifically toward enormous information stockpiling [3]. The thought of mapreduce may likewise be basically executed utilizing hadoop condition [4].Many procedures had been talked about in writing [5, 6, 7,8] to manage the disarranges of security in client server condition. However among the many controlled calculations utilized for security in administered situations Symmetric Encryption is more often than not anticipated because of its heartiness and ability [12] of usage in each 64-bit and 128-piece key arrangement. Inside the stylish situation because of the widen in the rate of utilization, upkeep of use, stockpiling of use, influenced the clients or manufacturers to attempt distributed computing environment. In this environment the product or information is put away chiefly inside the type of bunches. These groups will probably be transmitted over a cloud in light of the clients ask for assortments which will likewise be SAS, PAS and IAS [8]. Among the various offerings offered by method for the cloud climate, the conventionally utilized administrations include giving events on request and offering computational capacities on request.

The guide contract idea tended to in this paper helps the administered registering for extensive data units on groups of PC frameworks for offering figuring capacity on request. To encourage this bearer hadoop is conventionally utilized because of its ability of taking care of HDFS records in which information identified with elite machines close by the globe may likewise be spared. Map reduce is an execution of hadoop which helps in information preprocessing. This preprocessed information can be advantageous for the successful assessment of big data. Information mining is the investigation of learning with the motivation behind finding shrouded structure.

In some genuine capacities, it is basic to consider the difference in transient sides of a non-stationary time succession, and recognize the ones which can speak to the estimation of time circumstances. For instance, it is important in data spillage capacities from where the data has been spilled or it is intricate to recognize IP of an unapproved individual who logged whenever or sporadic interim of time in a cloud environment generally speaking such time arrangement are seen non-stationary. Customary time arrangement assessment employs statistical ways to deal with display and give a clarification to the information and predict future estimations of the time arrangement. It's not convenient, nonetheless, to decide the essential worldly examples of the time arrangement utilizing these ordinary methods. Utilizing a suite of perceptions, in this paper, we blessing a spic and span process for time arrangement data mining.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

By incorporating symmetric key encryption with the utilization of hadoop, temporal designs (client's log history at normal or irregular time interim) will likewise be easily distributed in non-stationary (cloud) air. Keeping in mind the end goal to handle the gigantic information and transmit the information all through the globe effective information change systems are to be adopted by methods for making utilization of symmetric encryption.

II. PREVIOUS WORK

P.SrinivasaRao et al[18] proposed a approach tolook after web usage from unauthorized clients with the aid of utilizinghadoopmapreduce where a namenode log file techniqueis proposed wherein identification of user's temporalpatterns process experimented.

ElisaBertino et al[10]proposed a process of Digital identification administration fora cloud using Multifactor Authenticationmanner .S.Fiseher-Hubnar et al[11, 15, 16] proposeda privacy and identification administration for Europe where itpresents privacy maintenance Authentication usingErroreneous Credentials.

Basker Prasad Rimal et al[8]. proposed a process in working out of taxonomy andsurvey of cloud computing methods Kumar Gunjan etal[13] gave an outline thought of identity administration incloud computing Mark D.Ryan et al[14, 17] explainedcloud computing protection: the scientific assignment andsurvey of options . Taking into consideration all of the abovedisorders, on this paper we are going to deal with defense ofcolossal data that has been transmitted in cloud throughHadoop disbursed process by means of making use of DES Algorithm.

To reduce the delay because of decryption process on thereceiving finish, an alternative procedure can also be adoptedfor authorized users by sending raw knowledge. In this paper we propose a novel methodology the place in the safety willbe offered in two phases in Hadoop Cloud atmosphere.

III. SYSTEM MODEL

In the Figure1 illustrated above, at any slave (datanode) user may send the data to any other data node in the cloud. The process of sending the data securely to other destination node is clearly visualized in the below Fig.1

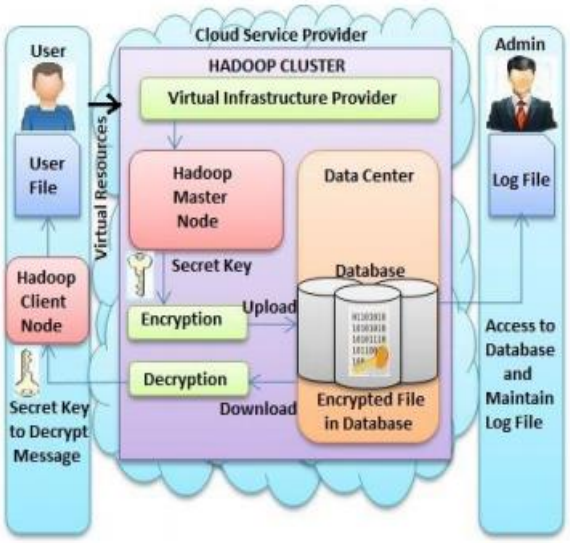

Fig. 1: Architecture of Hadoop Cluster in Cloud

At any datanode if the user is getting authorization to enter into cloud, he can be allowed to send or receive data that can be processed is shown in bellow Fig. 2.

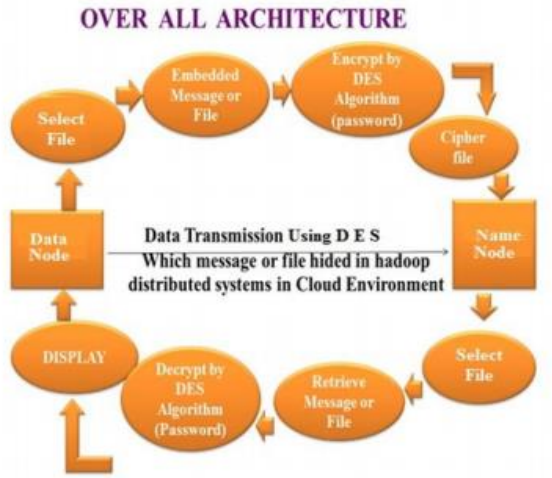

Fig. 2: Encrypted Data Transfer

A. Methodology

This paper tends to the strategy for relentless dataswitch to the validated clients and it contains amechanism where in the unapproved men and ladies may likewise beblocked from getting the data, and among the manyapproved people a security tag is connected so you could decide the wellspring of information spillages. The clients forwhom

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

the information is transmitted accepting to be authorized,if a data is released the security tag is prepared to 1 else thesecurity tag is prepared 0. For all these security labels wherethe hail is one, the comparing IP of the clients may bescrutinized and a mistake cautioning may be informed. In the event that themethod is rehashed the shopper or an individual correspondingto nitty gritty learning hub with the specific IP will beblocked from accepting extra data. With an end goal to avertunauthorized clients to see the substance material the information isencrypted influencing utilization of DES and File to key symmetricencryption calculations. Symmetric encryption which isused on this paper is additional pleasant than theuneven encryption which requires additional CPU cycles and CPU memory besides to a couple ofobstacles clarified by bundle File white et al [12].

B. Mapper and Reducer:

The mapper that includes my TPMap algorithm is applied to each input data that has been transmitted in hadoop allotted atmosphere .The information that is transmitted will have to be encrypted at each and every node with DESand Mapkey. The encrypted bundle of information will be stored at a customary memory of hadoop called HDFS(Hadoop allotted File approach). The patron called datanode is allowed to learn the understanding blocks if he is having authorization, otherwise tag worth will probably be incremented through quantity of times he attempted to grab the understanding. When the consumer at a distinctive knowledge node is having a tag price more than 0 can be recorded at log file of name node so that the writer will not be allowed to access any extra understanding in that cloud. That's the IP associated with that unique user shall be blocked. This processing shall be carried out within the Mapper and reducer whose job is to segregate all url's of distinctive consumer so that temporal patterns of the consumer can be located. A common algorithm with Map and cut back functions for determining such temporal patterns is listed within the table.1.

Table 1: TPMap and TPReduce Algorithm

Set the input path and the output path
Step 1: Client selects the file at any datanode.
Step 2: Authentication using Mapkey then goto step 3.
Step 3: Check for the captcha, if captcha is not matched and tag value is greater than 0 then goto Step 10 else goto step 4. // Start of Map Function.
Step 4: Map (key, value)
Step 5: Client can send or receive data
Step 6: Client encrypts file by using Encryption algorithm (DES) and Map key by giving Authentication (Password).
Step 7: The cipher file is transmitted over the cloud through hadoop HDFS.
// Start of Reduce Function
Step 8: Reduce (key, value).
Step 9: If at any IP, any unauthorized user is attempted to access the data, tag value will be set and the log file of name node will be updated so that the IP will be blocked.
Step 10: If the IP is with authorized user then decrypt data by giving password.
Step 11: The message or file is retrieved at any data node of the respective cloud.
Step 12: Logout from datanode.

C. Mapkey Algorithm:

The Mapkey algorithm is one of the securityalgorithms used to furnish safety for the userdataand store them in an encrypted format. A randomquantity is generated utilising Password based KeyDerivation function (PBKDF2) Algorithm that derivesthe important thing with the aid of making use of SHA1,SHA256,MD5 etc.,algorithms for the random generated number and againone-of-a-kind algorithms like AES,DES and so on., are applied throughsettling on the random quantity as a secret key. When aperson uploads the file in a cloud, the protection algorithmsare applied over the person file and encrypt the file andthen the cloud supplies an encrypted output file. Thesefiles are saved in a cloud storage database, which alsoprovide high degree safety to the cloud computingenvironment. A secret secret is supplied to the licenseduser to access his records in the cloud environment.

MapKey Algorithm:

1. Start
2. Read user file from at any datanode
3. Generate Random Number (n) // e.g.:12345
4. Perform PBKDF2 Algorithm to derive the Key (k)
5. Return MapKey
6. Stop

PBKDF2 Algorithm:

Input: *Pwd Password*
S salt Function
Ic Iteration Count

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

*Kl Key length in bits (2^{32} -1) * Hl*
Parameters: Prf → HMAC Function
Hl →Hash Function Digest System

***Output**: Mk →Master Key (Mk)*
***Algorithm:** if (Kl > (2 32 -1) *Hl)*
Return Error and Stop
Initialize L → [Kl / Hl]
*Q = Kl – (L-1) * Hl;*
For(i= 1 to l)
Xi=0;
V0=S || int(i);
For(j= 1 to Ic)
Vj= HMAC(Pwd,Vj-1);
Xi=Xi XOR Vj
Return Mk = Xi || X2 ||……. || Xl // <0…….Q-1>

As shown in the above TPMAP algorithm, the data encryption standards are one of the most protection algorithms used to furnish security to the person data and retailer them in an encrypted format. When a consumer uploads the file in a cloud, the protection algorithms are applied over the user file. These files are stored in a HDFS which is a normal hadoop Storage subject for all knowledge nodes in hadoop distributed atmosphere. A secret key is offered to the licensed users to access this data within the cloud environment.

IV. CONCLUSION

In this paper, a procedure is given to disclose thesafety structure to cloud environment. Thisframework helps in offering the security to the userknowledge in an encoded design that are transferred by utilizing theprovider client directly into a cloud, by incorpotating the keyfacets of unmistakable calculations like DES, FileKeymethods, that are set in a Hadoop group. Keyideas of this structure are the meaning of uniquesafety parameters for communicating securityrequirements and security execution, we alsoprovide a wellbeing way to deal with the cloud condition with the guide ofusing the security abilities and following the protectionparameters and security protection arrangements on the season of userlogin to give verification to the purchaser and to seek outtemporal patters inside the cloud.

REFERENCES

[1] KejaingYe et all, vHadoop: A Scalable hadoop virtual cluster platform for Map Reduce- Based Parallel Machine learning with Performance Consideration, 2012 IEEE International Conference on Cluster ComputingWorkships,PP:152-160, 2012

[2] J. Dean and S. Ghemawat, "Map Reduce:Simplified Data processing on large clusters",communications of the ACM.Vol.51, no.1,,pp107-113, 2008.

[3] T. White, Hadoop: The Definitive guide. yahoopress ,2010.

[4] Mohamed H. Almeer et al "cloud hadoopmapreduce for remote sensing image analysis"Journal of Emerging Trends in Computing andInformation Sciences, VOL. 3, NO. 4, April 2012.

[5] Abhipal Singh et al "File Transfer Using SecureSockets in Linux Environment", Proceedings ofthe 4th National Conference; INDIACom-2010.

[6] Matthieu Bloch et al "Network Security for ClientServer Architecture Using Wiretap Codes", IEEETRANSACTIONS ON INFORMATIONFORENSICS AND SECURITY, VOL. 3, NO. 3,SEPTEMBER 2008

[7] Xinyi Huang et al "Further Observations onSmart-Card-Based Password-Authenticated KeyAgreement in Distributed Systems", IEEETRANSACTIONS ON PARALLEL ANDDISTRIBUTED SYSTEMS, 2013.

[8] Bhaskar Prasad Rimal et al "A Taxonomy andsurvey of cloud computing system", Fifthinternational joint Conference on INC,I MS, IDC,2009.

[9] Poulami Dutta1 et al "Data Hiding in Audio Signal:A Review" International Journal of DatabaseTheory and Application Vol. 2, No. 2, June 2009

[10] Elisa Bertino et al, "privacy-preserving digitalidentity management for cloud computing" IEEEcomputer society technical committee on dataEngineering, 2009.

[11] S.Fischer-Hubner and H. Hebdon, "PRIMEprivacy and identity management for Europe"August 2010.

[12] kitu File whilte et al "symmetric vs Asymmetricencryption" 2010.

[13] Kumar Gunjan et al,international journal ofEngineering Research and Technology(IJERT),ISSN 2278-0181 vol1 issue4 june 2012.

[14] Make D.Ryan et al "Cloud computingSecurity:The Scientific journal of systems andSoftware challenge, and a survey ofsolutions",Elsevier,2013.

[15] Jittin et al . "An analysis on privacy preserving incloud computing", International journal ofcomputer trends and Technology(IJCTT),volume4 ,issue 6 june 2013.

[16] Man Qi* et al, "Social Networking searching andprivacy issues" information security technicalreport,Elsevier 2011