International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 14
October2016

A Novel Approach for Identification of Hadoop Cloud Environment

Chella Esther Varma, Associate Professor,
Vuppala Bhavana Eswar, Assistant Professor,
M.Sunil Kumar, Assistant Professor,
Geethanjali College Of Engineering And Technology,Cheeryal, R.R.Dist.

ABSTRACT: Due to the modern-day tendencies within the discipline of science and technology resulted in the trends of effective information switch, ability of handling colossal data and the retrieval of data efficiently. On account that the data that is saved is increasing voluminously, methods to retrieve relative understanding and protection associated concerns are to be addressed effectively to at ease this bulk data. Additionally with emerging standards of giant information, these security issues are a challenging undertaking. This paper addresses the limitation of comfortable data switch using the principles of data mining in cloud environment making use of hadoopmapreduce. Based on the experimentation achieved outcome are analyzed and represented with recognize to time and space complexity when compared hadoop with non hadoop approach.

KEYWORDS-Big Data, Hadoop, Mapreduce, Cloud Computing, Temporal Patterns

I. INTRODUCTION

The contemporary technological traits witnessed the storage of enormous data and methodologies special towards efficient retrievals. For the reason that this information is on hand are surmounting, safety breaches and upholding the privacy is a foremost trouble. These security issues are way more difficult whilst on the grounds that the data transfers in cloud atmosphere or parallel processing architectures [1]. So as to manage this data effectually ideas of mapreduce [2] is concentrated in the literature. This is as a result of its capabilities of faulttolarence and scalability in conjunction with simplicity.

One other fundamental advantage of highlighting the mapreduce notion is it allows the parallel processing environments which aid not directly in the direction of colossal data storage [3]. The notion of mapreduce may also be simply implemented utilisinghadoop environment [4].Many methodologies had been discussed in literature [5, 6, 7,8] to deal with the disorders of safety in clientr server environment. However among the many restrained algorithms used for safety in dispensed environments Symmetric Encryption is most of the time projected as a result of its robustness and capability [12] of utilization in each 64-bit and 128-bit key format. Within the trendy scenario as a result of the broaden in the rate of application, upkeep of application, storage of application, pressured the clients or manufacturers to undertake cloud computing atmosphere. In this atmosphere the software or data is stored principally within the form of clusters. These clusters will likely be transmitted over a cloud based on the clients request varieties which will also be SAS, PAS and IAS [8]. Among the many different offerings offered by way of the cloud atmosphere, the ordinarily used services comprise providing occasions on demand and offering computational capabilities on demand.

The map shrink concept addressed in this paper helps the dispensed computing for large data units on clusters of computer systems for offering computing ability on demand. To facilitate this carrier hadoop is ordinarily used due to its capacity of handling HDFS records in which data related to exclusive machines alongside the globe may also be saved. Mapreduce is a performance of hadoop which helps in data preprocessing. This preprocessed data can be worthwhile for the effective evaluation of bigdata. Data mining is the exploration of data with the purpose of discovering hidden structure.

In many real-world functions, it is fundamental to study the change of temporal sides of a nonstationary time sequence, and identify the ones which can be representing the value of time situations. For example, it is relevant in data leakage functions from where the data has been leaked or it is elaborate to identify IP of an unauthorized user who logged at any time or irregular interval of time in a cloud atmosphere as a rule such time series are viewed non-stationary. Traditional time sequence evaluation employsstatistical approaches to model and provide an explanation for the data andpredict future values of the time series. It's not convenient,nonetheless, to determine the primary temporal patterns of thetime series using these normal methods.Utilizing a suiteof observations, in this paper, we gift a brand new processfor time series data mining. By incorporatingsymmetric key encryption with the use of hadooop,temporal patterns (user's log history at regular orirregular time interval) will also be effortlessly published innon-stationary (cloud) atmosphere. In order to handlethe colossal data and transmit the data throughout the globeeffective data switch methodologies are to be adoptedby means of making use 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 data. 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 anyslave(datanode) user may send the data to any otherdatanode in the cloud. The process of sending the datasecurely to other destination node is clearly visualizedin the below Fig.1

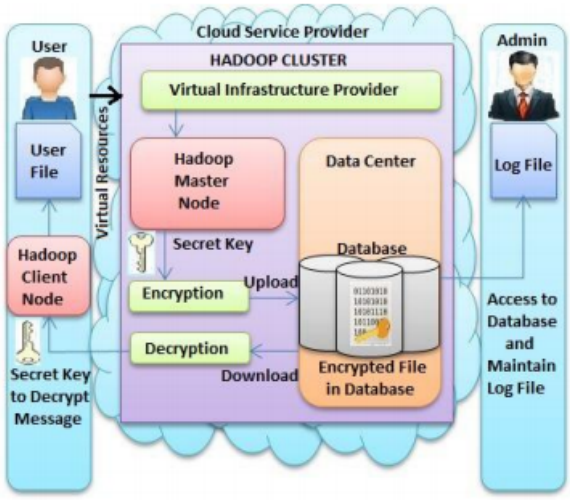

Fig. 1: Architecture of Hadoop Cluster in Cloud

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.

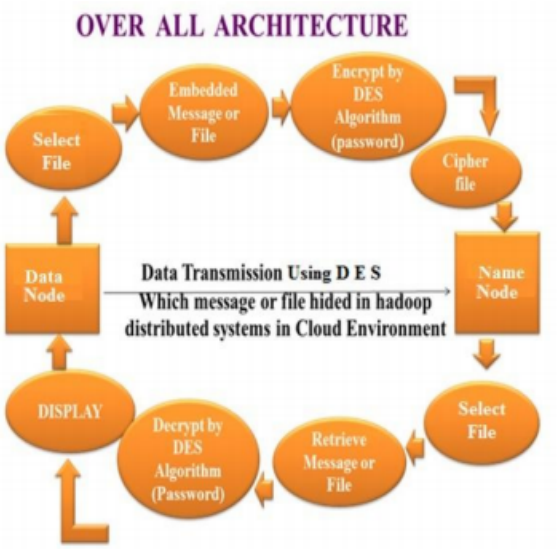

Fig. 2: Encrypted Data Transfer

A. Methodology

This paper addresses the method of mighty dataswitch to the authenticated clients and it comprises amechanism where in the unauthorized men and women may also beblocked from receiving the data, and among the manyapproved users a security tag is attached so that we coulddetermine the source of data leakages. The users forwhom the data is transmitted assuming to be authorized,if an data is leaked the security tag is ready to 1 else thesecurity tag is ready 0. For all these security tags wherethe flag is one, the corresponding IP of the clients might bescrutinized and an error warning might be notified. If themethod is repeated the user or a individual correspondingto detailed data node with the particular IP will beblocked from receiving additional data. In an effort to avertunauthorized users to view the content material the data isencrypted making use of DES and File key symmetricencryption algorithms. Symmetric encryption which isused on this paper is extra nice than theuneven encryption which requires extra CPU
cycles and CPU reminiscence moreover to a couple ofobstacles explained by package File white et al [12].

B. Mapper and Reducer:

The mapper that includes my TPMap algorithm isapplied to each input data that has been transmitted in
hadoop allotted atmosphere .The data that istransmitted will have to be encrypted at each and every node with DESand Mapkey. The encrypted bundle of data willbe stored at a customary memory of hadoop called HDFS(Hadoop allotted File system). The patron calleddatanode is allowed to learn the understanding blocks if heis having authorization, otherwise tag worthwill probably be incremented through quantity of times he attempted to grabthe understanding. When the user at a distinctive data nodeis having a tag price more than 0 can be recorded atlog file of namenode so that the writer will not be allowed toaccess any extra understanding in that cloud. That's theIP associated with that unique user shall be blocked.This processing shall be carried out within the Mapper andreducer whose job is to segregate all url's of adistinctive user so that temporal patterns of the usercanbe located. A common algorithmwith Map and cut backfunctions for determining such temporal patterns is listedwithin 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 againdifferent algorithms like AES,DES and so on., are applied throughsettling on the random quantity as a secret key. When auser uploads the file in a cloud, the protection algorithmsare applied over the user 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
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 dataencryption standards are one of the most protection algorithmsused to furnish security to the person data and retailer themin an encrypted format. When a consumer uploads the file ina cloud, the protection algorithms are applied over theuser file. These files are stored in a HDFS which is anormal hadoop Storage subject for all data nodes inhadoop distributed atmosphere. A secret key isoffered to the licensed users to access thisdata within the cloud environment.

IV. CONCLUSION

In this paper, an process is provided to explain thesafety framework for cloud atmosphere. Thisframework helps in offering the protection to the userdata in an encrypted layout that are uploaded by using theprovider user right into a cloud, by incorpotating the keyfacets of distinct algorithms like DES, FileKeymethods, that are placed in a Hadoop cluster. Keyideas of this structure are the definition of uniquesafety parameters for expressing securityrequirements and security performance, we alsoprovide a safety approach to the cloud environment with the aid ofusing the safety capabilities and following the protectionparameters and safety insurance policies on the time of userlogin to provide authentication to the user and to search outtemporal patters within the cloud.

REFERENCES

[1] KejaingYe et all, vHadoop: A Scalable hadoopvirtual cluster platform for MapReduce-

BasedParallel Machine learning with PerformanceConsideration, 2012 IEEE InternationalConference 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.