A Novel Approach for Security Issues Associated with Big Data

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

ABSTRACT: This paper introduces a designated analysis of between huge data and cloud computing safety disorders and challenges specializing in the cloud computing varieties and the carrier supply types. Big data is a data analysis methodology enabled through latest advances in applied sciences and structure. However, huge data entails a tremendous dedication of hardware and processing resources, making adoption costs of enormous data science prohibitive to small and medium sized companies. Cloud computing is a set of it offerings which are offered to a patron over a network on a leased groundwork and with the capability to scale up or down their carrier requisites. It benefits includes scalability, resilience, flexibility, effectivity and outsourcing non-core routine. It offers an progressive business model for corporations to adopt it offerings without upfront funding irrespective of the potential positive aspects performed from the cloud computing, the firms are slow in accepting it because of the security problems and associated challenges protection is among the major problems which bog down the development of cloud.

KEYWORDS-Cloud Computing, Big Data, Hadoop, Map Reduce, Hdfs (Hadoop Distributed File System).

I. INTRODUCTION

One of the vital prominent offerings supplied in cloud computing isthe cloud data storage, where subscribers don't need toretailer their data on their possess servers, where alternatively theirdata might be saved on the cloud service provider’s servers.In cloud computing, subscribers ought to pay the carriervendors for this storage service. This service does now notsimplest provides flexibility and scalability for the data storage,it additionally provide shoppers with the improvement of paying only forthe amount of data they ought to store for a unique periodof time, with none concerns for efficient storagemechanisms and maintainability disorders with huge amounts ofdata storage. In addition to those advantages, purchasers canconveniently entry their data from any geographical neighborhood wherethe Cloud provider servicer’s network or internet may also beaccessed. Data storage also redefines the safety issuesdistinct on user’s outsourced data (data that is notstored/retrieved from the users own servers). Sincecloud service providers (SP) are separate marketentities, data integrity and privacy are the most principal issuesthat need to be addressed in cloud computing.Furthermore, providing higher privateness as well as make sure dataavailability, can be carried out through dividing the data amonga few SP s in the market, founded on his to be hadprice range. Also we provide a determination for the client, towhich SP s he must chose to access data, with respect to dataentry satisfactory of carrier supplied through the SP s at the area ofdata retrieval.

In this survey we also furnish the user with higher assuranceof availability of data, by retaining redundancy in datadistribution. In this case, if a service provider suffers serviceoutage or goes bankrupt, the user still can access his data by means ofretrieving it from different service vendors. From the tradepoint of view, seeing that cloud data storage is a subscriptionprovider, the better the data redundancy, the bigger would be therate to be paid by means of the user.

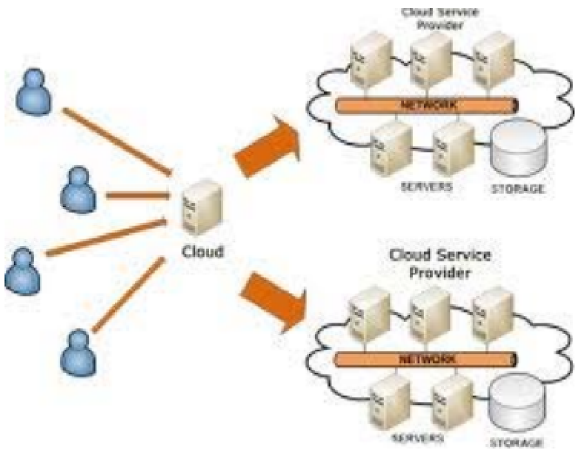

Fig.1: distribution of data over several SP's

Hence, we furnish an optimizationscheme to manage the tradeoff between the cost that a cloudcomputing user is inclined to pay to obtain a specified degreeof security for his data. In other phrases, we furnish a schemeto maximize the safety for a given budget for the clouddata.

This platform hides the complexity and important points of theunderlying infrastructure from customers and purposes with the aid ofproviding very simple graphical interface or API(applications Programming Interface) and also presents ondemand services which are invariably on, anywhere, every time andanywhere. It's a model for enabling easy, on-demandnetwork entry to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, andofferings) that may be swiftly provisioned and launched withminimal management effort or service provider intervention.Computing power and space for storing is supplied on-demandto organizations that outsource their IT management to the cloudservice provider. Cloud computing is a solution to increase thepotential or add capabilities dynamically without investing innew infrastructure, coaching new personnel, or licensing newprogram. However as increasingly understanding are placed within thecloud, considerations to grow about the safety of the cloudenvironment. Security problems in cloud computing has performed aprimary role in slowing down its acceptance. This work is asurvey more exact to the distinct protection problems and theassociated challenges that has emanated within the cloudcomputing process.

II. SECURITY ISSUES ASSOCIATED WITHDIFFERENT TOOLS

A. Big Data

Big data is a word used for description of massive amountsof data which are either structured, semi structured orunstructured. The data if it is not able to be handled by thetraditional databases and software tech ologies then wecategorize such data as big data the term big data is originatedfrom the web companies who used to handle looselystructured or unstructured data. The big data is defined usingthree v's.

 Volume: many factors contribute for the increase involume like storage of data, live streaming etc.

 Variety: various types of data are to be supported.

 Velocity: the speed at which the files are created andprocesses are carried out refers to the velocity.

Fig.1. Big data

Fig 1 suggests a natural computer virus data representation./ The areasfor illustration that is available in significant data are shown. Technologiesnow not simplest supports the collections of tremendous quantities such datacomfortably. Transactions which can be made far and wide the arena in abank, Wal-Mart purchaser transactions, and Face guide usersproducing social interaction data are few examples for significantdata usage.

B. HADOOP

This can be a freely available java based programmingframework supporting for the processing of giant sets of knowledgein a disbursed computing environment. Using Hadoop, enormousamount of data units may also be processed over cluster of serversand apps may be run on process with countless numbers of nodesinvolving terabytes of data as shown in Fig.2. Thislowers the chance of procedure failure even when a massive quantity ofnodes fail. It permits a scalable, flexible, fault tolerantcomputing answer. HDFS, a file procedure spanning all nodesin a Hadoop cluster for data storage links the file programs onneighborhood nodes to make it onto an extraordinarily tremendous file approach hencemaking improvements to the reliability.

□ Task trackers are accountable for strolling the tasksthat the job tracker assigns them

□ Job trackers has two main tasks whichare managing the cluster resources and schedulingall consumer jobs

□ Data engine contains the entire data in regards to theprocessing the data

□ Fetch manager helps to fetch the info at the same timedistinct challenge is running.

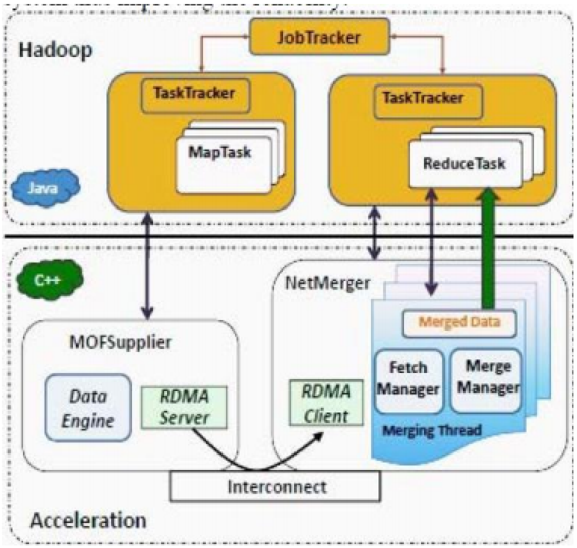

Fig.2. Hadoop structure.

C. Map Reduce

Map diminish framework is used to write down apps that system agiant amounts of data in a trustworthy and fault tolerant method asshown in Fig.3. The applying is at the beginning divided intocharacter chunks that are processed with the aid of user mapjobs in parallel. The output of map sorted by a framework andthen sent to the slash duties. The monitoring is taken care throughthe framework. The enter data is divided into userchunks and are offered for processing with the aid of the map assignment.These map task process the info in parallel and the outcomefrom the map undertaking is then offered to the decrease undertaking where the outcome which can be generated in parallel by means of the map project areconsolidated and the diminished report is given as output.

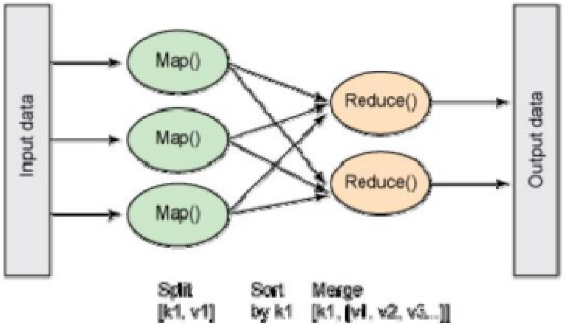

Fig.3. Map reduce.

Big Data Applications: In the present age of data explosion,parallel processing could be very much primary for performing alarge volume of data in a well timed manner. Parallelizationapproaches and algorithms are used to attain higherscalability and performance for processing massive data. Mapdecrease is an extraordinarily popularly used instrument or model used in industryand teachers. The two major advantages of map lessen areencapsulation of data storage, distribution, replication important points.

It is rather easy for use through the programmers to code for themap cut down challenge. Due to the fact the map reduces is schema free andindex free, it requires parsing of each file at the readingpoint. Map slash has bought plenty of attentiveness in thefields of knowledge mining, data retrieval, iamge retrievaland so on. The computation becomes problematic to be handled by means oftypical data processing which triggers the progress ofmassive data apps. Large data provides an infrastructure forretaining transparency in manufacturing industry, whichhas been having the ability to unreveal uncertainties thatexists in the element efficiency and availability.One more utility of the significant data is the field ofbio-informatics which requires giant scale data analysis.

III. THE PROPOSED APPROACHES

We present quite a lot of security measures which mighttoughen the safety of cloud computing environment. Given thatthe cloud atmosphere is a blend of many exceptionaltechnologies, we suggest quite a lot of options which togetherwill make the atmosphere convenient. The proposed solutionsinspire the use of more than one techanical/ tools to mitigatethe security hindrance unique in earlier sections. Protectionrecommendations are designed such that they don't lessenthe efficiency and scaling of cloud systems. Followingsafety measures will have to be taken to make sure the security in acloud atmosphere.

A. File Encryption: Since the data is present within the machines in a cluster, ahacker can steal all the valuable data. Accordingly, all of thedata stored will have to be encrypted. One of a kind encryption keysmust be used on specific machines and the important thing expertisemust be saved centrally at the back of powerful firewalls. This manner,even supposing a hacker is ready to get the data, he cannot extractsignificant data from it and misuse it. Userdata willbe stored securely in an encrypted method.

B. **File Encryption:** All the network communication will have to be encrypted asper enterprise specifications. The RPC system calls which takeplace will have to happen over SSL so that even though a hacker can tapinto network conversation packets, he are not able to extract pricelessunderstanding or manipulate packets.

C. Logging: All the map lower jobs which alter the info will have to belogged. Additionally, the understanding of users, which are liablefor those jobs, must be logged. These logs should beaudited most of the time to find if any, malicious operations arecarried out or any malicious user is manipulating the data inthe nodes.

Software format and Node maintenance: Nodes which run the program will have to be formatted almost alwaysto eliminate any virus present. All the software software'sand Hadoop software should be up-to-date to make the methodmore convenient.

E. Nodes Authentication: At any time when a node joins a cluster, it will have to be authenticated.In case of a malicious node, it should now not be allowed to become a member ofthe cluster. Authentication techniques like Kerberos may also beused to validate the approved nodes from malicious ones.

F. Rigorous system testing of Map scale back Jobs: After a developer writes a map minimize job, it should becompletely verified in a distributed environment alternatively of asingle laptop to be certain the robustness and stability of thejob.

G. Honey Pot Nodes: Honey pot nodes must be gift in the cluster, whichshow up like a normal node but is a lure. These honey pots trapthe hackers and integral movements would be taken to get rid ofhackers.

A layered framework for assuring cloud computingconsists of the secure virtual machine layer, secure cloudstorage layer, secure cloud data layer, and the secure virtualnetwork monitor layer. Cross cutting services are rendered bythe policy layer, the cloud monitoring layer, the reliabilitylayer and the risk analysis layer as shown in Fig.4.

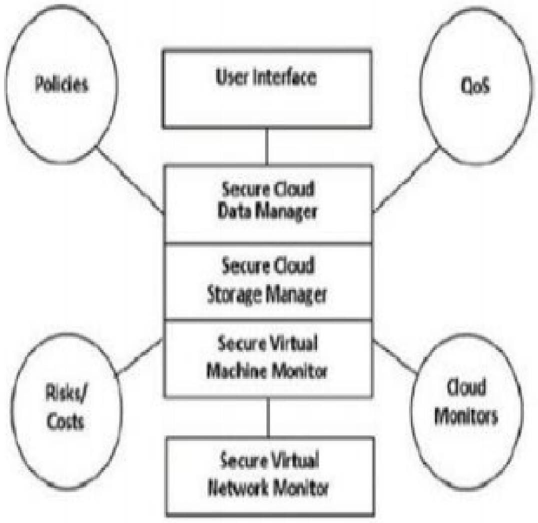
Policies
User Interface
QoS
Secure Cloud Data Manager
Secure Cloud Storage Manager
Secure Virtual Machine Monitor
Risks/ Costs
Cloud Monitors
Secure Virtual Network Monitor

Fig.4. Layered framework for assuring cloud.

Cloud computing helps in storing of data at a far flung websiteto be able to maximize resource utilization therefore, it is extremelyprimary for this data to be covered and access must begiven only to approved users. For this reason thisessentially amounts to relaxed third party authentication ofdata that is required for data outsourcing, as good as foroutside publications. In the cloud environment, the desktopserves the function of a third occasion publisher, which retailers thesensitive data within the cloud. This data wants to be protected,and the above mentioned systems ought to be applied tomake sure the maintenance of authenticity and completeness as shown in Fig.5.

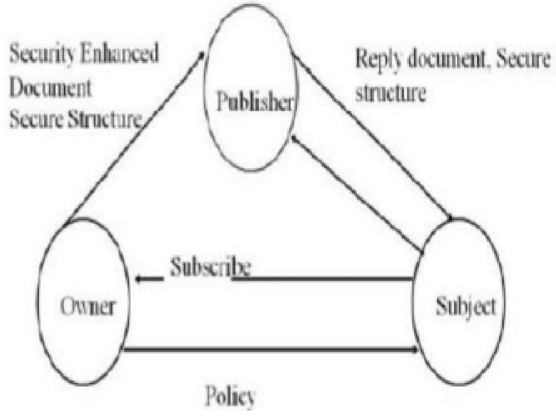
Security Enhanced Document Secure Structure
Publisher
Reply document, Secure structure
Subscribe
Owner
Subject
Policy

Fig.5. Third party secure data publication applied to cloud.

Integration of necessary entry control and differential privacy in dispensed atmosphere will be a excellent protectionmeasure. Data providers will manage the protection coverage oftheir touchy data. They will additionally manage the mathematicalsure on privateness violation that would take place. Within the aboveapproach, clients can participate in data computation without anyleakage of data. To prevent expertise leak, SELinux shall beused. SELinux is nothing but protection-enhanced Linux,which is a feature that presents the mechanism for helpingaccess manage security policy via the usage of Linuxsafety Modules (LSM) within the Linux Kernel. Enforcementof differential privacy can be carried out utilising amendment to Javavirtual desktop and the Map cut down framework. It will haveinbuilt applications which retailer the user identification pool for theentire cloud service. So the cloud provider will not ought tomaintain each and every user's identification for each and every software. Furthermoreto the above methodologies, cloud service will aid one third party authentication. The third party might be relied on via eachthe cloud provider and getting access to user. Third partyauthentication will add another safety layer to thecloud provider.

.

IV. CONCLUSION

This paper gave an outline of a scientific float ofsurvey of the tremendous data in the environment of cloud computing.We discussed in regards to the functions, benefits andchallenges faced through huge data when used over a cloudcomputing atmosphere. We proposed few solutions tolook after the data within the cloud computing atmosphere. Infuture, the challenges are ought to be overcome and make mannerfor the much more efficient use of the big data by the user on acloud computing atmosphere. It is rather a lot needed thatthe compuer scholars and IT professionals to cooperate andmake a successful and long run use of cloud computing andexplore new suggestions for the usage of the tremendous data over cloudenvironments.

REFERENCES

[1] Venkata Narasimha Inukollu, Sailaja Arsi, and SrinivasaRao Ravuri, “Security Issues Associated With Big Data InCloud Computing”, Vol.6, No.3, May 2014.

[2] Ren, Yulong, and Wen Tang. "A Service Integrity

Assurance Framework for Cloud Computing Based OnMapreduce."Proceedings of IEEE CCIS2012. Hangzhou:2012, pp 240 –244, Oct. 30 2012-Nov. 1 2012

[3] N, Gonzalez, Miers C, Redigolo F, Carvalho T, SimplicioM, de Sousa G.T, and Pourzandi M. "A Quantitative Analysisof Current Security Concerns and Solutions for CloudComputing.". Athens: 2011, pp 231 – 238, Nov. 29 2011-Dec. 1 2011.

[4] Hao, Chen, and Ying Qiao. "Research of CloudComputing based on the Hadoop platform.” Chengdu, China:2011, pp. 181 – 184, 21-23 Oct 2011.

[5] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C."Toward cloud computing reference architecture: Cloudservice management perspective.” Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.

[6] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues,challenges, tools and Good practices.” Noida: 2013, pp. 404 –409, 8-10 Aug. 2013.

[7] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Researchon Hadoop Cloud Computing Model and its Applications.”Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.

[8] Wie, Jiang, Ravi V.T, and Agrawal G. "A Map-ReduceSystem with an Alternate API for Multi-core Environments.”Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[9] K, Chitharanjan, and Kala Karun A. "A review onHadoop HDFS infrastructure extensions.” JeJu Island: 2013,pp. 132-137, 11-12 Apr. 2013.

[10] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C."Business model canvas perspective on big dataapplications." Big Data, 2013 IEEE International Conference,Silicon Valley, CA, Oct 6-9, 2013, pp. 32 - 37.

[11] Zhao, Yaxiong, and Jie Wu. "Dache: A data awarecaching for big-data applications using the MapReduceframework." INFOCOM, 2013 Proceedings IEEE, Turin, Apr14-19, 2013, pp. 35 - 39.

[12] Xu-bin, LI, JIANG Wen-rui, JIANG Yi, ZOU Quan"Hadoop Applications in Bioinformatics." Open CirrusSummit (OCS), 2012 Seventh, Beijing, Jun 19-20, 2012, pp.48 - 52.