IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Secured Hadoop as A Service Based on Infrastructure Cloud Computing Environment

Atul U. Patil[1], R.U.Patil[2], A.P.Pande[3], B.S.Patil[4]

Assistant Professor, PVPIT, Budhgaon[1,3]

Assistant Professor, BVCOE, Kolhapur[2]

Associate Professor, PVPIT, Budhgaon[4]

Abstract: This paper propose system that gives out requirement based allocation of Hadoop as a service with infrastructure cloud environments. Hadoop is effective big data analysing and processing platform these days. But in instead of its promising nature, researchers or professional organizations are not technically sound or having capacity to implement and maintain a working Hadoop environment. That's, we are providing the secured Hadoop as a service. For cloud services to use computing services or analytics services by the cloud users is truly problematic. It's a big issue to complete user's needs. Hence. On-Demand Hadoop service through cloud infrastructure provides a way to handle big data on the go. Potentially strengthening in security problems and achieves equal Job scheduling and quick process of huge information in less quantity of time and resources by computing the scientific or any high performance computing jobs. Hadoop and Cloud taken the apps and software systems and also different databases to cloud data centres, wherever the handling of the sensitive data and processes is not safe. Its security loophole. Solution is given by processing and securing the information using ciphering deciphering and storing them into the cloud servers.

Keywords: Hadoop as a Service, Encryption/decryption algorithm, Storage utilization.

I. INTRODUCTION

Cloud considered as a quickly rising new technology for delivering computing as a utility. In cloud computing varied cloud customers demand type of services as per their dynamically ever-changing needs. Thus it's the work of cloud to avail all the services to the consumers. But as a result of the supply of limited number of resources it's very troublesome for cloud CSP to produce all services. From the cloud providers' perspective cloud resources should be allotted in a very honest manner. So, its important thing to complete cloud customers' Quality requirement regarding services. So for accessibility a provider has to extra provide keep an outsized proportion of nodes idle so CSP unable to complete needs of customers. The necessity to stay of these nodes idle results in low utilization. Way improve to keep least secondary servers ideal. But this occurs to mostly rejecting a tremendous amount of requests to some extent at that a provider now not provides on-demand computing [2].

Other consideration when evaluating Hadoop provider of service with which the can accomplish elastic demand. Anyone can think how openly Hadoop as service can accomplish changing demands for compute and storage resources with ease without worrying about hectic of managing and implementing Hadoop. For example, Hadoop Big data jobs produce lots of intermediate results that may be temporarily stored and on cloud secondary servers. Hadoop as a service transparently expand and contract storage without system administrator intervention. Hadoop administrators not required to fix parameters or risk delaying jobs. Also show how well the Hadoop as a service from cloud handles workloads. Organizations that process all high performance jobs and scientific analysis by science industry will face a large traffic of mixture of workloads. In the past decades Infrastructure Service in cloud has become an attractive service to the provision and management of computing resources. An important thing of Infrastructure service cloud is providing customers on-demand control to computing infra. So to provide use based resources, CSP should either keep extra resources on (or pay a high value for operative resources underutilized) or remove an oversized proportion of user requests (in that case the access isn't any longer on-demand). At the same time, not all users need really need basis access to infrastructure [3].

Several Jobs and scientific workings are based on opportunistic systems wherever pauses in service are possible. Here a system is proposed, Hadoop as a service with the help of infrastructure service that gives out on-need based providing the computing infrastructure.

The target is to handles and process big data in less CPU clock cycles and keep users away from hassles of configuring Hadoop at remote servers and strengthening the use of resources by executing jobs through fair4s algorithm for equal distribution of jobs, additionally increase the utilization of CPU by introducing upload/download activity. Data security kept intact through RSA algorithm also used DES for comparison .And used one which not only increase security but also gives better utilization.

They give various techniques for analysing, processing and modelling workloads. However, the job properties execution policies are many in those systems from the ones in a Hadoop system.

Copyright to IJARCCE

DOI 10.17148/IJARCCE.2016.53261

II. THE PROPOSED SYSTEM

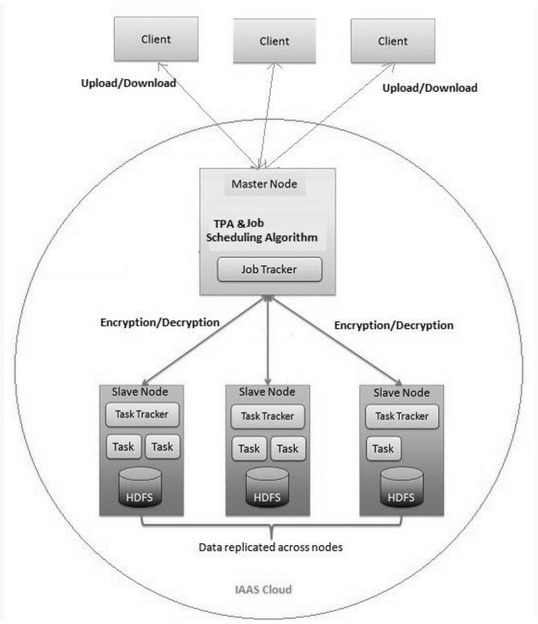

Fig.1 IAAS Cloud Architecture

Hadoop Cloud provides solution for CPU, storage, time parameters of improvement however moving massive amounts of knowledge in associated in cloud presented an insurmountable challenge [4].Cloud computing is a very undefeated paradigm of service destined computing and has revolutionized the means computing infrastructure is abstracted and used. Following most well-liked cloud Solutions include:

1. Infrastructure Cloud Services
2. Platform Cloud Services
3. Software as Cloud Services

Thought can even be extended to Storage as a Service. Various database services also provided through cloud on air. Changes in information access patterns of application and therefore the have to be compelled to scale intent on thousands of commodity Hardware led to birth of a replacement category of systems referred to as Key-Value stores [11].Area of big data analytics, we propose the Map Reduce paradigm as Hadoop as a service through cloud with open-source implementation Hadoop, in terms of usability and performance.

The System implements these modules:

1. Hadoop Cloud Configuration
2. Hadoop As Service Login portal
3. Hadoop Cloud Administrator portal
4. Job Scheduling Algorithm
5. Encryption/decryption module
6. Third Party Auditing.

3.1 Hadoop Configuration (Hadoop as a Service)

The Hadoop is a framework that permits for the decentralized process of big data across clusters of computers using straightforward programming models which is map reduce model. it's designed to proportion from single to then thousand servers, providing massive computation and storage capacity, instead of think about underlying hardware to give large availability, the infrastructure itself is intended to handle problems, thus delivering a most available service on prime of a cluster of nodes, every of which can be vulnerable to failures [6]. Hadoop implements Map reduce, using HDFS. The Hadoop Distributed File System gives clients to possess namespace, unfold across several lots of or thousands of clusters, making one big file system.

Framework allows to process large data with ease. Any of those splits (also told fragments or blocks) may be execute on secondary servers within the cloud infrastructure. The present Hadoop system consists of the Hadoop architecture, Map-Reduce, the Hadoop distributed file system.

JobTracker is routine for allocating and running MapReduce jobs in Hadoop on master server node. There's one Job tracker executes on hadoop cluster. Job tracker works JVM. And slave node is assigned with task tracker node location. JobTracker in Hadoop performs; scheduling of assignments to task trackers [9].

A TaskTracker is slave node service within the Hadoop secondary nodes that takes and process tasks like Map, reduce operations from a JobTracker. Single task tracker per node. Task tracker runs on its JVM. Each TaskTracker is having vacant slots, these tells the amount of jobs that it will settle for. The TaskTracker has JVM to workout tasks this is often to confirm that process failure doesn't take down the task tracker [10].

The Hadoop Distributed File System (HDFS)
HDFS is a fault tolerant and self-healing distributed filing system designed to point out a cluster of business normal servers into a massively scalable pool of storage. Developed specifically for large-scale process workloads where quality, flexibility and turnout square measure necessary, HDFS accepts data in any format despite schema, optimizes for prime system of measurement streaming, and scales to tried deployments of 100PB and on the way side [8].

3.2 Hadoop as Service Login and Registration
It offered Interface to Login. Client will login to the Hadoop as a service through user interface and then upload the file and download file from web application to hadoop cloud where we have already configured the Hadoop processing which will process the data through efficient scheduling like fair4s algorithm and all obtain the detailed summery of his account. During this means security is provided to the consumer by authentication information for logging in to portal and stores it in info at the most server that ensures the safety. Client activities records kept and used for audit trails. With this facility, it ensures enough security to consumer and information hold on at the cloud servers solely may be changed by the consumer.

3.3 Hadoop Cloud Administrator
It is administration of Hadoop Services and Cloud infrastructure. Cloud service supplier has an authority to feature and take away clients and configure Hadoop

DOI 10.17148/IJARCCE.2016.53261

services. It ensures enough security on client's information hold on at the cloud servers. Conjointly the log records of every registered and authorize consumer on cloud solely will access the services. This specific consumer log record is helps in improve security.

3.4 Fair4s Algorithm

Algorithm processing any kind of workload small, large jobs Users specify the workload in terms of a map, reduce operations Programs written during this purposeful style area unit Automatically parallelized and executed on an oversized cluster of commodity machines. [7].

Our implementation of Fair4s algorithm runs on an oversized cluster of commodity machines and is very scalable. Map-Reduce is Popularized by open-source Hadoop project. Our fair4s algorithm works on process of enormous files by dividing them on variety of chunks and assignment the tasks to the cluster nodes in hadoop multimode configuration. In these ways in which our planned Fair4s Job scheduling algorithm improves the utilization of the Cloud secondary servers with parameters like time, CPU, and storage. Various features of the Job scheduling algorithm are enlisted below.

3.4.1 Features of Fair4s:

Extra functionalities available in Fair4s algorithm create it workload efficient than efficient measure listed out below these functionalities permits algorithm to provides out efficient performance in process huge work load from totally different clients.

1. Setting Slots Quota for Pools- All jobs are divided into many pools. Every job belongs to at least one of those pools. Whereas in Fair4S, every pool is designed with a maximum slot occupancy. All jobs belonging to a uniform pool share the slots quota, and also the range of slots employed by these jobs at a time is restricted to the utmost slots occupancy of their pool. The slot occupancy higher limit of user teams makes the slots assignment a lot of versatile and adjustable, and ensures the slots occupancy isolation across totally different user teams. Though some slots are occupied by some giant jobs, the influence is barely restricted to the native pool within.

2. Setting Slot Quota for Individual Users-In Fair4S, every user is designed with a most slots occupance. Given a user, regardless of what number jobs he/she submits, the entire range of occupied slots won't exceed the quota. This constraint on individual user avoids that a user submit too many roles and these jobs occupy too several slots.

3. Assigning Slots based on Pool Weight- Fair4S, every pool is designed with a weight. All pools that look ahead to a lot of slots type a queue of pools. Given a pool, the prevalence times within the queue is linear to the burden of the pool. Therefore, high waited pool are allotted with a lot of slots. Because pool weight can be changes and so small job fairness comes into picture.

4. Extending Job Priorities- Fair4S introduces an in depth and quantified priority for every job. The task priority is described by associate degree integral range ranged from zero to a thousand. Generally, at intervals a pool, a job with a better priority will preempt the slots used by another job with a lower priority. A quantified job priority contributes to differentiate the priorities of small jobs in numerous user-groups. Programming Model

3.4.2 Fair4s Algorithm

Fair4S, which is modeled to be fair for small jobs. In variety of working situations tiny jobs are huge and lots of them require instant responses, which is an important factor at production Hadoop systems. The ineffective nature of hadoop schedulers and GFS read write algorithm for working with tiny sized jobs motivates us to use and analyze Fair4S, which introduces pool weights and extends job priorities to guarantee the rapid responses for small jobs [1] In this scenario clients is going to submit jobs through client login on master server where the Fair4s executes. On master server the Auditing functions and equal distribution of job is done through our proposed algorithm in efficient manner.

3.4.3 Procedure of Slots Allocation

1. The primary step is to allot slots to job pools. Every job pool is organized with two parameters of maximum slots quota and pool weight. In any case, the count of slots allotted to a job pool wouldn't exceed its most slots quota. If slots demand for one job pool varies, the utmost slots quota is manually adjusted by Hadoop operators. If a task pool request for extra slots, and decision taken by checking quota and wait for slot allocation. The scheduler allocates the slots by round-robin algorithm. Probabilistically, a pool with high allocation weight are additional likely to be allotted with slots.

2. The second step is to allot slots to individual jobs. Every job is organized with a parameter of job priority that may be a worth between zero and a thousand. The duty priority and deficit are removed and mixed into a weight of the duty. Inside employment pool, idle slots are allotted to the roles with the highest weight.

3.5 Encryption/decryption

In this, data get encrypted/decrypted by exploitation the RSA encryption/decryption algorithm encryption/decryption algorithm uses public key & map, private key for the encryption and decipherment of data. Here we have tested different Encryption/Decryption algorithm and the performance in terms of CPU Utilization, Storage , and Time is very good of RSA Algorithm as compared to other algorithms and also provides the in depth security Consumer transfer the file in conjunction with some secrete/public key so private key's generated &amp; file get encrypted. At the download time using the public key/private key pair expected job decrypted and downloaded. Like client upload the file with the public key and also the file name that is used to come up with the distinctive private key's used for encrypting the file. During this approach uploaded file get encrypted and store at main servers and so this file get splitted by using the Fair4s Scheduling algorithm that provides distinctive security feature for cloud data. In an exceedingly reverse method of downloading the data from

DOI 10.17148/IJARCCE.2016.53261

cloud servers, file name and public key wont to generate secrete and combines

The all parts of file so data get decrypted and downloaded that ensures the tremendous quantity of security to cloud information.

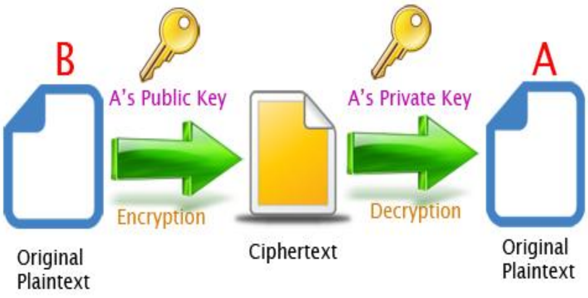

Fig.2 RSA encryption/decryption

3.6 Administration of client files(Third Party Auditor)
This module provides facility for auditing all client files, as numerous activities are done by client. Files Log records and got created and hold on Main Server. for every registered client Log record is get created that records the varied activities like that operations (upload/download) performed by client. Additionally Log records keep track of your time and date at that varied activities carried out by client.

For the security and security of the client data and conjointly for the auditing functions the Log records helps. Additionally for the Administrator Log record facility is provided that records the Log info of all the registered clients. In order that Administrator will control over the all the info hold on Cloud servers. Administrator will see client wise Log records that helps us to notice the fraud information access if any fake user attempt to access the info hold on Cloud servers.

III. RESULTS

Results of this underlying project will be explained well with the help of project work done on number of clients and one Master server and then five to ten Slave servers so then taken results bases on following parameters taken into consideration like

1) Time

2) CPU Utilization

3) Storage Utilization.

Our evaluation examines Provided RSA Encryption/Decryption algorithm provides better performance on cloud infrastructure as compared to the DES algorithm in Storage, Time, and CPU also get improved broadly. In this ways we have not only provided the Hadoop as a service through the cloud but we also taken into consideration the Security aspect and provided secured hadoop as a service on infra clouds.

Uploading and Encrypting Data Using RSA and DES Algorithm. Result are as below

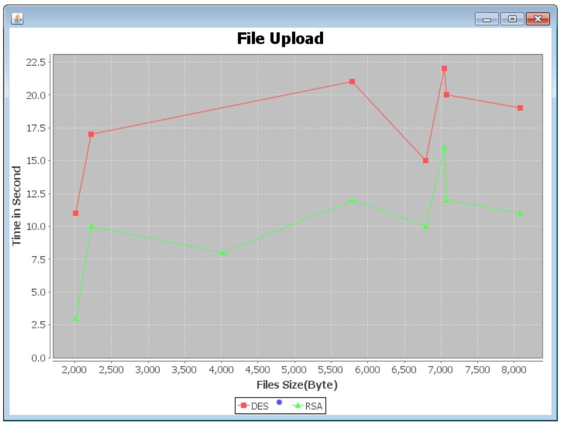

Fig 3 Results Encryption/Decryption Algorithms.

Decrypting and Downloading Data Using RSA and DES Algorithm. Result are as below

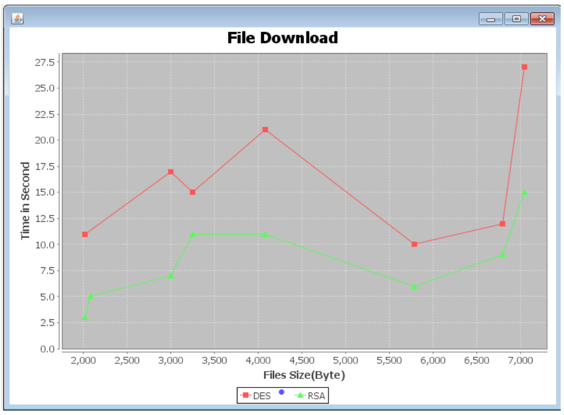

Fig 4 Results Encryption/Decryption Algorithms.

CPU Utilization Result are as below

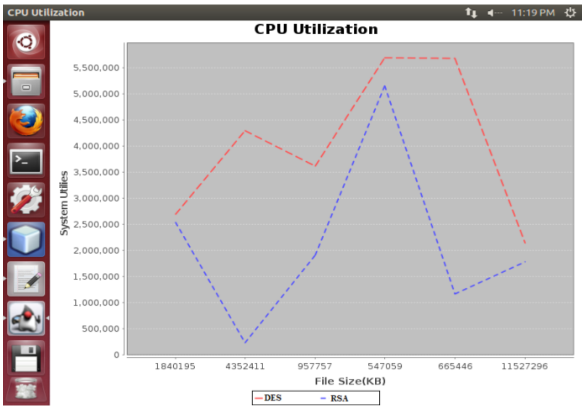

Fig 5 CPU Utilization RSA Vs DES

IV. CONCLUSION

We have proposed secured Hadoop as a service cloud service that provides need based of Hadoop service through infrastructure cloud with optimized utilization, opportunistic provisioning of cycles from idle nodes to different processes with resolving the data security issues through use of efficient encryption/decryption algorithm. But in instead of its promising nature, not all companies or professional organizations are technically sound or capable of implementing and maintaining a successful Hadoop environment. Because of its difficult nature to manage. As a result, we are providing the secured Hadoop service. Which keeps end user away from all these hassles and keep processing and analysing big data throughout the industry Hence all unutilized nodes that remains idle are all get utilised because of hadoop map reduce nature and mostly improvement in security problems and achieves load balancing and quick process of huge data in less amount of your time. For file uploading and file downloading; and optimizes the processor utilization and storage space use. During this paper, we tend to additionally plan a number of the techniques that area unit implemented to guard data and propose design to protect data in cloud. This model was proposed to store data in cloud in encrypted information using RSA technique that relies on encryption and decryption of data. Till currently in several planned works, there's Hadoop configuration for cloud infrastructure. However still the cloud nodes remains idle.

Thus Secured Hadoop as a service on the infrastructure clouds will provide the way to process big data on cloud with provided security to the data and along with the hassle free platform which keeps away users from Hadoop configurations and gives out Hadoop service as web application service.

REFERENCES

[1] I. Raicu, I. Foster, and Y. Zhao. Many-Task Computing for Grids and Supercomputers. In Many-Task Computing on Grids and Supercomputers, 2008. MTAGS 2008. Workshop on, pages 1–11, Nov.2008.

[2] Open Crowd. Cloud Taxonomy, Landscape, Evolution. http:/ / www. opencrowd.com/assets/images/views/views_cloud-taxlrg. png. June 8, 2010.

[3] T. Zhang, W. Lin, Y. Wang, S. Deng, C. Shi, and L. Chen,“ The design of information security protection framework to support smart grid,” in Conference on Power System Technology, 2010, pp. 1 –

[4] Nivyakant Sgrawal et al., “Big Data and Cloud: Current State and Future Opportunities”, EDBT, pp 22-24, March 2011.

[5] B. KRISHNA KISHORE UPPE NANAJI AND Y.SWAPNA ‘‘Improving Utilization of Infrastructure Clouds, ISSN: 2278-7844

[6] Rofrey Cean et al., “Simplified data processing on Big clusters”, communications of the acm, Vol S1, No. 1, pp.107-113, 2008 January.

[7] ENISA. Benefits, risks and recommendations for information security. European Network and Information Security Agency. November 20, 2009.

[8] D. Mavulya, J. Jan, R. Gandhi, and P. Narasimhan, ‘an Analysis of Traces from a Mapreduce Cluster,’’ in Proc. CGGRID, 2010, pp. 94-103.

[9] K. Mean et al.,“ a flexible data processing tool”, In CDCM, Jan 2010.

[10] T. Rtonefraker et al., “MapReduce and parallel DBMSs: friends or foes?” In CACM. Jan 2010.

[11] S. Das et al., “Ricardo: Integrating R and Hadoop”, In SIGMOD 2010.

[12] J. Cohen et al.,“MAD Skills: New Analysis Practices for Big Data”, In VLDB, 2009.

[13] Srizhen Rang et al., “ SaaS-Based Cloud Computing in the University Research and Teaching Platform”, ISIEE, pp. 230-243, 2013.

[14] AmazonWeb Services LLC. Amazon Simple Storage Service. http: //aws.amazon.com/s3/, 2009.

[15] H. Tang, R. Zin, S.A. Brandt, E.L. Miller, D.D.E. Long, and T.T. Mclarty, ‘‘File System Analytics forLarge Scale Scientific Computing Applications,’’ in Proc. MSST, 2004,

[16] Yu, Jie, Guangming Liu, Wei Hu, Wenrui Dong, and Weiwei Zhang. "Mechanisms of Optimizing MapReduce Framework on High Performance Computer" 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, 2013.

[17] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, “The cost of doing science on the cloud: the montage example,” in Proceedings of the 2008 ACM/IEEE conference on Supercomputing. IEEE Press,2008, pp. 1–12.

[18] J. Sztipanovits, J. A. Stankovic, and D. E. Corman, “Industry Academy Collaboration in Cyber Physical Systems (CPS) Research,” CRA, Tech. Rep., 2009..

DOI 10.17148/IJARCCE.2016.53261