Deduplication & Optimize Task Scheduling in Cloud

Machala Rama Krishna,
M-Tech Student, RITM, Bengaluru,

Mr. Mylara Reddy,
Assistant Professor, RITM, Bengaluru.

ABSTRACT-As we know cloud provides services like Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Software as a Service (SaaS).Under these category STORAGE outsource is one of the we known service from the cloud. Individual person and entities can access the software and hardware such as network, storage, server and applications which are located remotely easily with the help of Cloud Service. The tasks/jobs submitted to this cloud environment needs to be executed on time using the resources available so as to achieve proper resource utilization, efficiency and lesser makespan which in turn requires efficient task scheduling algorithm for proper task allocation. In this project, we have introduced deduplication with encrypted data management. Dynamic Proof of Storage (PoS) is a useful cryptographic primitive that enables a user to check the integrity of outsourced files and to efficiently update the files in a cloud server. Time Scheduler is used for optimized task scheduling.

INTRODUCTION

Storage outsourcing is becoming more and more attractive to both industry and academia due to the advantages of low cost, high accessibility, and easy sharing. Storage out sourcing works on the basis of pay as you use concept, users no need to invest for the storage infrastructure. Many companies, such as Amazon, Google, and Microsoft, provide their own cloud storage services, where users can upload their files to the servers, access them from various devices, and share them with the others. Just storing information or files to the cloud will not provide security assurance for the end users. Deduplication with encryption technique is used to avoid duplicate data in cloud storage.

A new type of computer model i.e. Cloud provides the user to access the applications and associated data from anywhere. Cloud computing which is a huge distributed computing environment contains a large amount of virtualized computing resources available for individual or an organization. Major challenge of introducing cloud is to provide the guarantee of Quality of Service (QoS) which is quite challenging. For effective resource utilization scheduling of jobs and maintaining load is required in cloud storage system. This can be achieved by executing tasks in both primary and secondary node. By considering parameters such as scalability, resource utilization, cost, computational time, priority, performance, bandwidth, resource availability and many more for optimum task scheduling.

Literature Survey

Title: Survey of various scheduling algorithm in cloud computing environment.
Author: Pinal Salot

This paper illustrates the 3 main algorithms used in Cloud environment:
Max-min,
Min-min and
RASA

Each of these algorithms estimate the completion and execution time of each submitted task on each available resource. In Max-Min algorithm task which is taking maximum time to complete the task will schedule first and task which is taking minimum time will execute last so minimum task has wait for completion of all task those are taking maximum time. In case of Min-Min job scheduling algorithm task which is taking minimum time will execute first and task which is taking maximum time will execute at the end. Drawback of these two algorithm is overcome by Resource Aware Scheduling Algorithm (RASA).In RASA job scheduling algorithm, execution time and completion time will calculate based on available resources and then RASA will alternatively allocate Max-Min and Min-Min algorithm to overcome drawback of these 2 algorithms. RASA is a hybrid algorithm of both Max-Min and Min-Min job scheduling algorithms.

Title: Review on Max-Min Task Scheduling Algorithm in Cloud Computing.
Author: Bhavisha Kanani and Bhumi Maniyar

Cloud Computing/service allows user to use the computing, software, platforms and infrastructure as a services via internet. For using these resources customer no need to invest money on computing infrastructure instead of that pay only for what they use. In this project survey author is done on Max-Min task scheduling algorithm and drawback of this algorithm. The primary objective of this paper is optimizing the scheduling polices. In this paper task scheduling is explained how jobs are submitted to cloud environment onto available resources in such a way that the total response time, and makespan (time required for completing the task) is minimized. Cloud computing system offers virtual machine which is scalable but job scheduling to these virtual machine is a major problem.

Title: Load Balancing in Cloud System using Max-min and Min-min Algorithm
Author: Rajwinder Kaur and Pawan Luthra

This paper explains the types of load balancing algorithm as below:

Batch mode heuristic scheduling algorithms (BMHA): In this mode first it will collect all task in batch and then it will schedule jobs/tasks based on below mentioned fashion/logic type:

FCFS – First Come First Serve,

RR – Round Robin fashion,

Max-Min and

Min-Min

Online mode heuristic algorithms: In this type task will not shored in batch mode instead of that jobs are scheduled as and when they arrive.

Following metrics are focused for Load Balancing Algorithms in Cloud:

Scalability: Load balancing algorithm should work continue even after adding some nodes into the cloud network.

Resource Utilizations: Algorithm should effectively use the resources available in cloud system.

Response Time: Load balancing algorithm should quickly response to the task and also it should take minimum response time.

Fault Tolerance: Author also explained load balancing algorithm should take care of any node failure in cloud system.

Overhead Associated: Total amount of overhead required for implementing load balancing algorithm and it should be minimized for better performance.

In this paper literature survey is done how the load balancing algorithm distributes the work to all available nodes such that no nodes should be free and no node should be overloaded.

Title: Deduplication on Encrypted Big Data in Cloud

Author: Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng.

In this paper author proposed following schemes:

Data ownership challenge: Challenges faced by data owner while uploading file to cloud system.

Proxy Re-Encryption (PRE) to manage encrypted data storage with deduplication: In this scheme regeneration key will be provide to the data holder if he upload the same content to cloud system. It will avoid duplicate data and provide key for re-encryption such that data holder/ subsequent user will get decryption key for encrypted original data.

Symmetric encryption to deduplicate encrypted data: In this scheme it will not only avoid the storage of redundant copy on top of that it will encrypt the content by using symmetric encryption and then it will store into cloud system so that it will provide secure storage system. This scheme also support

Following features are proposed in this paper:

Data Ownership Verification: Before uploading file to cloud system user ownership verification is required.

Data Deduplication: Data deduplication is achieved by avoiding not to store the same content into the cloud system. It will also manage the encrypted data by using AES encryption and RSA key generation.

Data Deletion: If any data delete request comes from data holder this scheme will check is it a original single copy if so it will permanently delete the content. If that storage is shared by other users then it will delete the record of having key to encrypt the storage for that particular user and it will not give access to this user.

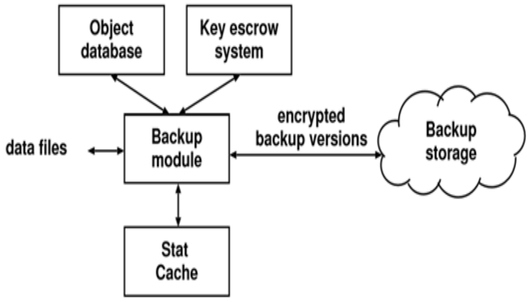
Encrypted Data Update: Provide support for how to update encrypted data in cloud system.

Title: A secure cloud backup system with assured deletion and version control

Author: H.Chen

Cloud storage is an popular service that provides individuals and entities to outsource the storage of data backups to remote cloud providers at a low cost without investing money for infrastructure. Cloud user looking for security guarantees of their outsourced data backups in cloud system. In this paper author proposed Fade Version, which is a secure cloud backup with secured encrypted data.

Author explained architecture of Fade Version as depicted below:



This paper provides following features:

Provide security guarantees for user’s outsourced data

Support of assured deletion, for example, data files are permanently inaccessible upon requests of deletion for that particular user.

Version control support for user outsourced data backups, so that cloud users can roll-back to extract data from previous versions of stored data.

This paper lists the following cloud backup systems in the market available:

Dropbox,

Jungle Disk,

Nasuni

Title: Boosting efficiency and security in proof of ownership for deduplication.

Author: A.Sorniotti

In this paper author explained storage deduplication with secured system. Author is focusing on following 2 concepts:

Proof of Ownership (POW)

Secured deduplication

In storage outsource cloud system just maintaining of deduplication without security for the outsourced data will not suffice the user requirement. Now a days user is looking for secured system for their outsourced data in cloud system. Hence author proposed deduplication with encrypted data management.

Implementation of this concept introduces many security risks. In this paper author address the proof of ownership which required for data upload and download to the cloud. Secondly author provided secured deduplication data outsource system. At last the quality of proposed schemes are supported by extensive benchmarking.

Existing System

Existing system avoids the duplicate data storage and stores only single copy, But for the stored copy these is no security. This scheme works in single use environment. This system will not satisfy the end user, since user wants his outsourced data should secured. Disadvantage of this system is it is not considered encryption technique while file uploading in cloud storage system. In cloud storage system many data outsource requirements will come dynamically, So handling of multiple requirement or task required task scheduling technique in order to use available resource effectively. Apart from task scheduling system also find difficulties in node overload or idle problem. Since security is not provided for the storage system any third party can access this user data and he can use this data for his benefit or it may used for misuse. Existing system is fail to achieve in proper resource utilization, efficiency and lesser makespan which in turn requires efficient task scheduling algorithm for proper task allocation.

Proposed System

Proposed system will provide following features by overcome the disadvantages of existing system as explained in the above section:

Storage outsource by managing encrypted data with deduplication.

Avoid network traffic by scheduling the job during non-peak hours.

Provides data security by encryption & decryption while uploading/downloading data from cloud.

Load balancing by executing tasks in both primary and secondary node.

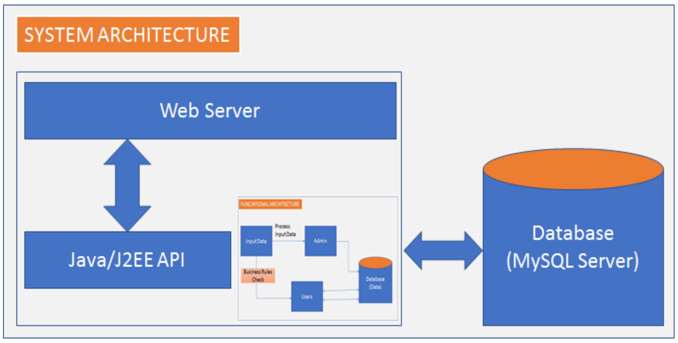
Task scheduling and easy monitoring by automated time scheduler.

Load balancing should be achieved based on the system storage load on the specific cloud server.

Security issues will be handled by using efficient algorithms maybe AES, RSA for encryption and decryption process.

System Architecture:

Following diagram illustrate the system architecture for proposed system.



User interacts to the cloud server for uploading his file. Java/J2EE API's MySQL is used for project implementation.

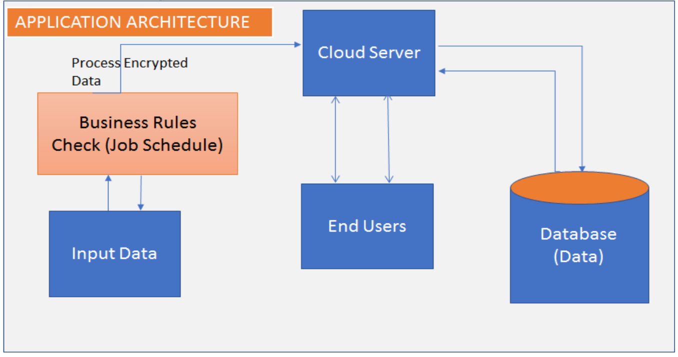
Application Architecture:

Below diagram illustrate the application architecture with how the business rules are performed which includes:

Deduplication with encrypted data handling before providing data outsource service to the end user.

Time scheduling.

Load balancing.



Need For Project

To avoid duplicate data and save the storage space in cloud.

Avoid network traffic by scheduling the job during non-peak hours.

Security encryption & decryption needs to be implemented using AES and RSA Key algorithm.

Task scheduling and easy monitoring by automated time scheduler.

Motivation

Cloud Technology.

Increase performance key comparison.

Secure way of data storage.

Avoid duplicate data to be stored in cloud which is more costly.

System Configuration

We have following software and hardware requirements to implement our proposed energy-efficient routing mechanism based on the cluster-based method for the mobile sink in WSNs with obstacles and evaluate it's performance.

HARDWARE REQUIREMENTS:

System	:	Pentium IV 2.4 GHz.
Hard Disk	**:**	500 GB.
RAM	**:**	4 GB

Any desktop / Laptop system with above configuration or higher level.

SOFTWARE REQUIREMENTS:

Operating system	:	Windows XP / 7/8
Language used	:	Java (JDK 1.8)
Web Technology	:	Servlet, JSP
Front End	:	HTML, CSS, JavaScript
Web Server	:	Tomcat 8.0
IDE	:	Eclipse
Database	:	My-SQL 5.5

Conclusion

Efficient data storage outsource is provided by using data deduplication with secured data storage technique in cloud system. For data encryption Advanced Encrypted Storage algorithm is used. Key required for symmetric encryption RSA algorithm is used. Keys are shared to the user by e-Mail notification. Proposed implementation has proved to achieve high cost savings. Explored this project and found that it will reduce up to 90-95% storage required for backup related applications and up to 68% in standard file systems.

Job scheduling algorithm like Max-min and Min-min are applicable in small scale distributed systems. Limitation of these two algorithms are taken care in RASA algorithm by alternatively allocating these two algorithm with available resource. In computational cloud environment, high throughput and load balancing is equally important. Our proposed system provides load balancing by executing tasks in both primary and secondary nodes in such a way that neither nodes are idle nor nodes are overloaded. The further enhancement of this work can be done by providing higher level security with SHA-2 technique. As of now keys are sharing through e-mail it can be further enhanced by sharing keys through SMS.

Acknowledgements

I am grateful to my Reva Institute of Technology and Management, for providing me with the facilities for successfully completing II- phase project evolution project work “Deduplication and Optimized task scheduling in cloud” and providing me all the required inputs and feedbacks for successful completion of this paper. I will take this opportunity to thanks to my guide Mr Mylara Reddy for his valuable guidance. I will also thank our college management team for their continued support. Finally, I thank my wife and children for their motivation, moral, and material support.

References:

[1] Bhavisha Kanani and Bhumi Maniyar, “Review on Max-Min Task Scheduling Algorithm in Cloud Computing,” Journal of Emerging Technologies and Innovative Research, Volume 2, Issue 3, March 2015.

[2] Yash P. Dave, Avani S. Shelat, Dhara S. Patel and Rutvij H. Jhaveri, “Various Job Scheduling Algorithms in Cloud Computing: A Survey,” ICICES 2014 - S.A.Engineering College, Chennai, Tamil Nadu, India, ISBN No. 978-1-4799-3834-6/14, IEEE 2014.

[3] Rajwinder Kaur and Pawan Luthra, "Load Balancing in Cloud System using Max-min and Min-min Algorithm," International Journal of Computer Applications (0975-8887), NCETCT-2014.

[4] Deduplication on Encrypted Big Data in Cloud Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE.

[5] A secure cloud backup system with assured deletion and version control- H. Chen.

[6] Survey of various scheduling algorithm in cloud computing environment - Pinal Salot

[7] R.D. Pietro and A. Sorniotti, "Boosting efficiency and security in proof of ownership for deduplication," Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, pp. 81-82, 2012, doi:10.1145/2414456.2414504.

[8] Dropbox, “A File-Storage and Sharing Service,” http://www.dropbox.com/.

[9] Google Drive, http://drive.google.com.

[10] Mozy, “Mozy: A File-storage and Sharing Service,” http://mozy.com/.