CROSS CLOUD MAPREDUCE FOR BIGDATA

Pragya Jaju

Abstract- The dramatic growth of data volume in recent years imposes an emerging issue of processing and analyzing a massive amount of data. As a prominent framework for big data analytics, Map Reduce plays a crucial role. We look at a geo distributed cloud architecture in this research that provides Map Reduce services based on large data acquired from end customers all over the world. Existing work handles Map Reduce workloads using a classic computation-centric strategy, in which all input data from many clouds is aggregated to a single location. Existing work handles Map Reduce workloads using a classic computation-centric method in which all incoming data from many clouds is consolidated into a single virtual cluster. We propose a unique data-centric architecture with three main techniques: cross-cloud virtual cluster, data-centric job placement, and network coding based traffic routing, due to its low efficiency and high cost for large data support. Our concept yields an optimization framework for operating a series of Map Reduce operations in dispersed clouds with the goal of minimising both computation and transmission costs. We also create a parallel algorithm by breaking down the original large-scale problem into numerous distributively solvable subproblems that are coordinated by a higher-level master problem.Finally, we undertake real-world experiments and comprehensive simulations to demonstrate that our idea beats previous work significantly.

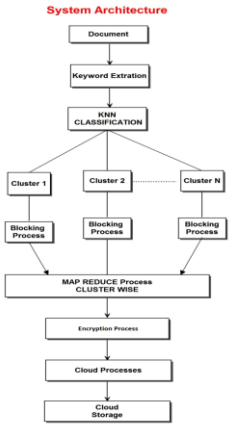
Keywords- cloud, deduplication, mapreduce

I. INTRODUCTION

A vast number of businesses use Map Reduce to parallelize their data processing on distributed computing systems. It breaks down a job into a series of parallel map tasks, followed by reduce tasks that combine all of the map tasks' intermediate results to produce the final results. Map Reduce jobs are typically run on commodity PC clusters, which necessitate a significant investment in hardware and maintenance. A cluster is underutilized on average because it must be supplied for peak consumption to avoid overload. As a result of its flexibility and pay-as-you-go business model, cloud becomes an attractive platform for MapReduce jobs.

II. ARCHITECTURE

We investigate a distributed cloud architecture with many clouds in various geographical locations. It provides a platform for worldwide applications that collects data from end users all over the world and delivers a set of services on that data, such as searching, sorting, and data mining. A direct graph G(N,A) can be used to model this dispersed cloud system, where N and A signify cloud locations and dedicated inter-cloud linkages, respectively. Each cloud provides infrastructure for both storage and computing. The data acquired from the respective regions is constantly stored in the storage clouds. The compute cloud is made up of a network of virtualized and networked servers. The computation cloud contains a collection of interconnected and virtualized servers. The stored input data are organized as multiple blocks, Given a set of MapReduce jobs V of different types, each job $v \in V$ is assigned a virtual cluster.

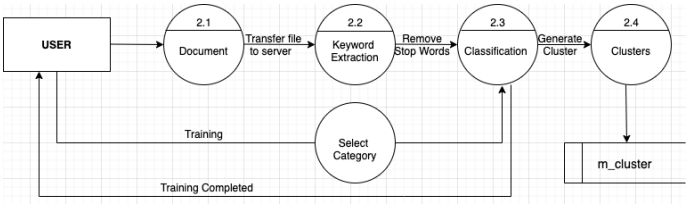


III. IMPLEMENTATION

1. Training and classification

This module is responsible for training the model based on KNN classification algorithm. It takes a set of keywords for training and creating a cluster for each category. After the cluster is created, the algorithm scans through the uploaded file to remove the auxiliary verb, conjunctions and other such stop words.

2. Block Splitting and Merging Technique



When dealing with the issue of transporting documents through an organisation or transferring to the web when the record size is large, the square parting and merging approach is used. Parting the record will solve the problem in certain cases. The records will be divided into small lumps, which will be converged into a single document at the goal. While the file is uploaded it is split into blocks of equal sizes and while downloading they are merged to a single file.

3. Logical Block Addressing Technique

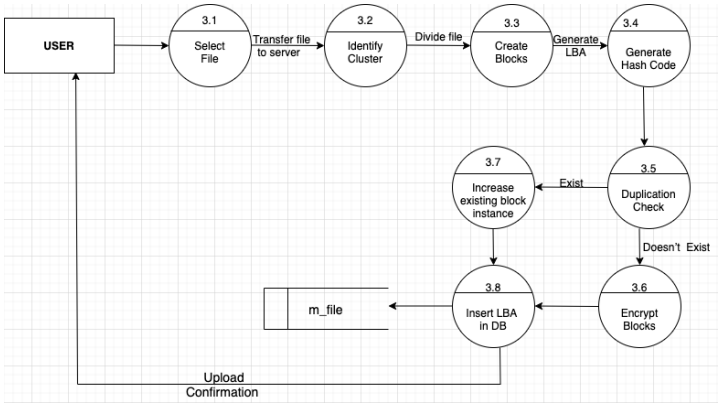
When a client uploads a document, it will be divided into multiple chunks. For each chunk of data a number will be generated based on the previous block and stored, similarly while downloading LBAs are fetched for a particular file and merged to get a single downloadable file.

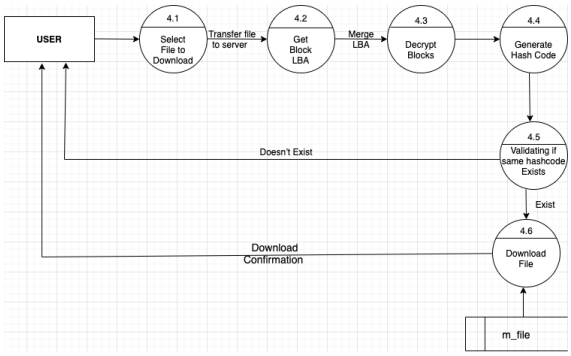
4. Cryptographic Technique

Cryptographic techniques are used to ensure secrecy and integrity of data in the presence of an adversary. It is used when a file is uploaded to encrypt the blocks of the file and decrypt them before downloading.

5. Block Existence Verification Technique

When a client uploads a file, it will be divided into many blocks, each with a different hash value. To download a file, hash code is generated again inorder to match with the existing ones and if the match is not found, it means that the data has been altered and the block no more belongs to the same file hence it will not be downloaded. However, if the hash values matches, a request is sent to the server and using logical block addressing it finds the block numbers of the file, merges and decrypts them to further download on the client machine.





IV. CONCLUSION

The MapReduce for Big data using Cloud was successfully implemented and tested. The system used a KNN Classification to train the cluster for specific category. As for every project, this project comes with some limitations which were not possible to achieve in the provided time-frame. There is also a scope for improvement and modification in the future. Based on the limitations in the first build of the system, one can jot down possible methods or techniques or algorithms to overcome them in the next build. This will ensure that the product can be made commercially standard and of market quality.

V. REFERENCES

[1]. Shunrong Jiangy, Tao Jiang and Liangmin Wang, 'Secure and Efficient Cloud Data Deduplication with Ownership Management', (1939-1374) 2017 IEEE

[2]. Ran Ding, Hong Zhong, Jianfeng Ma, Ximeng Liu, and Jianting Ning,'Lightweight Privacy-Preserving Identity-Based Verifiable IoT-Based Health Storage System', (2327-4662) 2019 IEEE

[3]. Shuguang Zhang, Hequin Xian, Zenpeg Li and Liming Wang, 'SecDedup: Secure Encrypted Data Deduplication with Dynamic Ownership Updating', 2020, IEEE Access

[4]. Wenting Shen, Ye Su, Rong Hao,' Lightweight Cloud Storage Auditing With Deduplication Supporting Strong Privacy Protection', 2020, IEEE Access

[5]. Xinrui Ge, Jia Yu, Hanlin Zhang, Chengyu Hu, Zengpeng Li, Zhan Qin, Rong Hao, 'Towards Achieving Keyword Search over Dynamic Encrypted Cloud Data with Symmetric-Key Based Verification' , IEEE Transactions on Dependable And Secure Computing, vol. , no. , 2018 1

[6]. Wei Guo, Hua Zhang, Sujuan Qin, Fei Gao, Zhengping Jin, Wenmin Li, Qiaoyan Wen, 'Outsourced dynamic provable data possession with batch update for secure cloud storage' , 2019 Elsevier B.V.

[7]. Yongkai Fan, Xiaodong Lin, Gang Tan, Yuqing Zhang, Wei Dong, Jing Lei, 'One secure data integrity verification scheme for cloud storage' , 2019 Elsevier B.V.

[8]. Hui Cui, Robert H. Deng, Yingjiu Li, and Guowei Wu, 'Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud', Journal of Latex Class Files, vol. , no. , Month 2016

[9]. Giuseppe Ateniese, Randal Burns, Reza Curtmola, Joseph Herring, Lea Kissner , Zachary Peterson, Dawn Song, 'One secure data integrity verification scheme for cloud storage', 14th ACM Conference on Computer and Communications Security(CCS 2007)

[10]. Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart, 'Message-Locked Encryption and Secure Deduplication' ,Eurocrypt, 2013