

A Similarity based Compression Approach for Efficient Data Processing on Cost Optimized Multi-Cloud

S.T. Deepa (deepatheodore@gmail.com)¹, L. Gejalakshmi²

¹ Department of Computer Science, Shri S.S.Shasun Jain College, Chennai, India.

² Chennai Higher Secondary School, Chennai, India.

Abstract

Cloud computing provides a promising platform for flexible massive storage of data, computing and software services in a scalable manner. Massive storage sensing is prevalent in both industry and research applications where the data storage consumes volume and high velocity. There are five phases in the compression based cost effective multi cloud architecture.

Pre-process phase: In this phase user is intended to upload their local files. The cloud server decides whether these files to be uploaded considering the authenticity of the user and the content priority.

Deduplication process: Once the file is approved for uploading, similarity model is used for compression and clustering, Data chunks are created. Similarity model works with text data and multidimensional numerical data is contributes to the majority of the data available. Markov model is used to calculate the similarity in text data and in tree topology, similarity is determined by the number of leaf nodes. The data is checked for duplication in this step. It is done by generating signature for the data chunk, the signature stored in DDB (Deduplication DataBase) is compared with the signature of the chunk to be stored and then data chunk is stored.

Upload phase: After approval for uploading and clustering, the data chunks are encrypted and uploaded into multi-cloud where the application and data are fragmented and stored in multcloud to enhance security and protection. In multi cloud architecture, no cloud provider learns the complete application logic and overall application results which leads to data and application confidentiality.

Update phase: If user intend to modify, insert or delete some blocks of the existing files, then the corresponding data chunks alone is updated in the cloud.

Proof of storage Phase: The user has a meta data file stored locally to identify which data chunks is available in which cloud in our multi cloud architecture.

To make cloud storage cost effective, on demand resource provisioning is established and the cost is calculated depending on the number of user and number of resources used. On demand resource planning avoids under and over provisioning of the resource enhancing the resource utilization.

Keywords: *Multicloud, Data compression, Resource provisioning*