DATA DEDUPLICATION IN CLOUD

CLOUD COMPUTING AND VIRTUALIZATION -(ITE3007)

REVIEW 1 PROJECT REPORT

Submitted by

ARIHANT JAIN

20BIT0006

SAHIL NANDAL

20BIT0010

RAJ RATNESH SINGH

20BIT0028

Submitted to

Dr Siva Rama Krishnan S

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

VIT UNIVERSITY: VELLORE 632 014

JANUARY, 2023

CONTENTS

1. **Abstract**
2. **Introduction**
3. **Expected Outcomes**
4. **References**

ABSTRACT

Cloud storage is one of the cloud computing services that provides consumers with virtualized storage on demand. However, as the amount of data in the cloud grows, customers expect to be able to use on-demand cloud services at any time, while providers must maintain system availability and handle a significant amount of data. As a result, removing data redundancy has become one of the most important approaches in the cloud. Providers require a method to drastically reduce data quantities in order to save money while running large-scale systems. One of the reasons for the increase in the cost of cloud hosting services can be the bandwidth pricing on the usage-based whereas in the pay as you go service model helps in decreasing the bandwidth usage cost. One of the points of attention is the bandwidth cost which is a serious topic of concern. Research on Traffic Redundancy Elimination (TRE) [1] technologies have been done to decrease the bandwidth cost while data transfer. TRE technologies eliminate the transmission of duplicate information. TRE has garnered a lot of attention in the cloud environment due to its excellent rate of reduction in bandwidth cost. One of the ways for reducing data redundancy is to implement sender-based TRE or receiver-based TRE but it cannot capture all types of traffic redundancy as traffic can be either short-term (time span of seconds or minutes) or long-term (time span can hours or days). Moreover, the receiver based TRE solution is inefficient when the data changes occur which creates a need for a new method which is more efficient and the solution proposed for that is Collaborative end-to-end (CoRE) [2].

INTRODUCTION:

Background (System Study Details in brief):The advent of the 20th century brought an immense increase in digital computation and, along with it, the start of an information era. This has further led to the development of research to ensure easily accessible data, which led to the introduction of cloud computing as a basic building block for the same. Cloud computing is a synthesis of numerous computing domains that allows for lower capital expenditures and greater flexibility in resource supply, scaling up, and scaling down. This is where cloud storage plays a major role in providing any kind of service. Because of benefits and characteristics such as multi-tenancy, improved server utilization, energy efficiency, and elasticity obtained from on-demand utility computing services, cloud computing use has increased tremendously. Data deduplication is a type of data compression technique that eliminates multiple copies of repeated data. It is also known as intelligent compression or singleinstance storage. This approach is used to enhance storage utilization and can also be used to reduce the amount of bytes that must be delivered during network data transfers. During the deduplication process, unique blocks of data, or byte patterns, are discovered and stored. Deduplication techniques use data similarity to locate duplicate data and save storage space. Data deduplication is so closely related to incremental backup, which transfers just the data that has changed since the previous backup.

EXPECTED OUTCOMES

The development of a system that transmits data from the transmitter to the receiver. At the receiver, the TCP data stream that is entering is divided into pieces. The chunks are placed in a local chunk storage after being joined together in a chain. The receiver compares each arriving piece to the chunk storage. When a matching chunk is located on a chain, the following chunks are retrieved as predicted chunks for incoming data in the future. In a PRED message, the sender is informed of the signatures and anticipated offsets in the incoming data stream for the retrieved chunks as a forecast for the sender's impending outgoing data.

REFERENCES

[1] **Traffic Redundancy and Elimination approach to Reduce cloud Bandwidth and Costs**

Ushanandini Balu, Resmi S

Available - https://www.irjet.net/archives/V2/i3/Irjet-v2i3297.pdf

[2] **CoRE: Cooperative End-to-End Traffic Redundancy Elimination for Reducing Cloud Bandwidth Cost**

Lei Yu, Haiying Shen, Karan Sapra, Lin Ye

Available - https://www.researchgate.net/publication/303890468_CoRE_Cooperative_End-to-End_Traffic_Redundancy_Elimination_for_Reducing_Cloud_Bandwidth_Cost