The Office of Technology Management

UNIVERSITY OF TEXAS ARLINGTON

RapidCDC: A Method of Content-Defined Chunking for Rapid Deduplication

Tech ID: UTA 19-12

INVENTORS: Song Jiang, Fan Ni

TECHNOLOGY NEED

Deduplication is a technique to remove redundant data in a data storage system to reduce demand on storage space. Content defined chunking (CDC) is a commonly used deduplication method which involves a number of steps, including partitioning each file into chunks of variable sizes based on content, hashing to obtain unique fingerprints of individual chunks, and comparison of the fingerprints to remove the redundant content. In this method, the first step about file chunking is one of the most expensive operations. In this step, the chunk boundaries are usually determined by using a rolling window approach, which checks the file content byte-by-byte. This makes CDC deduplication much more expensive and time-consuming than fixed-size chunking (FSC), which determines chunk boundaries at fixed file positions to produce chunks of constant size, though CDC can have a much higher deduplication ratio.

INVENTION DESCRIPTION/SOLUTION

We have developed a modified CDC algorithm known as RapidCDC to improve the chunking speed without reducing the deduplication ratio. The RapidCDC performs chunking and fingerprint generation in the same method as traditional CDC, but the process of determining the chunk boundaries is not performed byte-by-byte. Instead, an accelerated method of detecting the chunk boundaries is used. Here the boundary of the first duplicate chunk is found using the rolling window, after this the chunk boundaries in a file can be determined at the same speed as fixed-size chunking method until a non-duplicate chunk is encountered. The accelerated chunking speed is available as long as there exists a sequence of duplicate chunks, which is common in various data sets. With accelerated chunking speed, the chunking time is reduced almost to zero.

APPLICATIONS

- Deduplication technique to improve data storage system efficiency.

KEY BENEFITS

- Increase in chunking speed up to 40 times faster, or almost eliminate the CDC chunking cost.
- Its performance is positively correlated to deduplication ratio.
- Its deduplication ratio is the same as the existing CDC approach.
- Its implementation in an existing CDC deduplication system does not require any major change of its operation flow.

STAGE OF DEVELOPMENT

Prototype

Extensive tests done

INTELLECTUAL PROPERTY STATUS

Provisional

About the Inventor:
Song Jiang

Contact information
For licensing, please contact
Koffi Selom Egbeto
(Licensing Associate)
koffi.egbeto@uta.edu
otm@uta.edu
P: 817.272.1132

Our mailing Address:
The Office of Technology Management
701 S Nedderman drive,
Suite 350, Arlington, TX 76019

Connect with us: