ABSTRACT

Data deduplication, an efficient approach to data reduction, has gained increasing attention and popularity in large-scale storage systems due to the explosive growth of digital data Data deduplication is a lossless compression technology that has been widely used in storage systems for space optimization.

Deduplication has become a widely deployed technology in cloud data centers to improve IT resources efficiency. However, traditional techniques face a great challenge in big data deduplication to strike a sensible tradeoff between the conflicting goals of scalable deduplication throughput and high duplicate elimination ratio. *AppDedupe*, an application-aware scalable inline distributed deduplication framework in cloud environment, to meet this challenge by exploiting application awareness, data similarity and locality to optimize distributed deduplication with inter-node two-tiered data routing and intra-node application-aware deduplication. It first dispenses application data at file level with an application-aware routing to keep application locality, then assigns similar application data to the same storage node at the super-chunk granularity using a handprinting-based stateful data routing scheme to maintain high global deduplication efficiency, meanwhile balances the workload across nodes.

. AppDedupe builds application-aware similarity indices with super-chunk handprints to speed up the intra-node deduplication process with high efficiency. The experimental evaluation of AppDedupe against state-of-the-art, driven by real-world datasets, demonstrates that AppDedupe achieves the highest global deduplication efficiency with a higher global deduplication effectiveness than the high-overhead and poorly scalable traditional scheme, but at an overhead only slightly higher than that of the scalable but low duplicate-elimination-ratio approaches.

Index Terms: big data deduplication, application awareness, data routing, handprinting, similarity index