druva

Tech brief

Cost Efficient Backups For The Enterprise

Deduplication Done Right with Druva

Q120 20048

Overview

Enterprise data is growing at a phenomenal rate and it's located everywhere: on endpoints, in data centers and increasingly in the cloud. Industry studies indicate that 90% of enterprises are now using one or more cloud services to house 65% of their business-critical workloads.[1] Using cloud-based services also opens up new use cases including cloud-based backup, disaster recovery and data intelligence.

It has become harder for organizations to manage and protect their data across increasing volumes and data types. Data deduplication is one of the technologies changing data protection economics resulting in the technology becoming a critical component to any organization.[2]

This brief explains the different types of deduplication technology as well as why Druva's source global deduplication delivers significant value to any data protection strategy.

Basics of Deduplication

There are three attributes associated with data deduplication: location, logic and scale.

Location refers to where the deduplication process takes place. It can be at the target or source side

Logic defines the granular level of deduplication. Granularity can be file level (two different files), subfile level (block level), and application aware where mapping technology is used to dedupe the application

Scale is the need to dedupe across multiple devices. For example, the CEO has sent out an email to all employees resulting in the organization needing to dedupe across each user's device

Types of Deduplication

Target Deduplication

Target deduplication requires that full or incremental backups be transmitted across the network inside a backup container. Deduplication occurs once the data arrives at the target system, usually a hardware appliance. Benefits of this technology include: transparent integration with existing backup software and no impact on the source for the deduplication process. Unfortunately, this approach does little to reduce the impact on the network between the source and target environment and scaling occurs by purchasing additional appliances. For a typical 100 TB file, customers need to purchase 150-250 TB appliance. Once the 100 TB backup data file is deduped and a second 100 TB file comes along, you need a second hardware appliance. Duplicate data is created as data cannot be deduped across multiple appliances.

Source Deduplication

Source deduplication is completed at the very beginning of the backup process. With source deduplication, the entire data stream is sliced into shards or chunks, and hashes are calculated for each chunk/shard. The hash is sent to the dedupe server, where the "hash lookup" process takes place. If the hash lookup determines the hash has never been seen before, the data is seen as new and unique, and it is transferred across the network to the target. If the hash has been seen before, the hash table records another instance of the same chunk/shard. The advantage of source dedupe is that it reduces network traffic as well as storage requirements. The con for source deduplication is the need for greater computational power on the source to compute the hash as well as integration with libraries to complete the deduplication process.

---

1 Taneja Group, 2017
2 eBizQ: The Business Benefits of Deduplication

It should be noted that this additional compute power is minimal compared to the network impact of passing full and incremental backups to a dedupe appliance.

Table 1 below provides a summary of the benefits and costs for both deduplication types.

Table 1. Target and Source Deduplication Pros and Cons

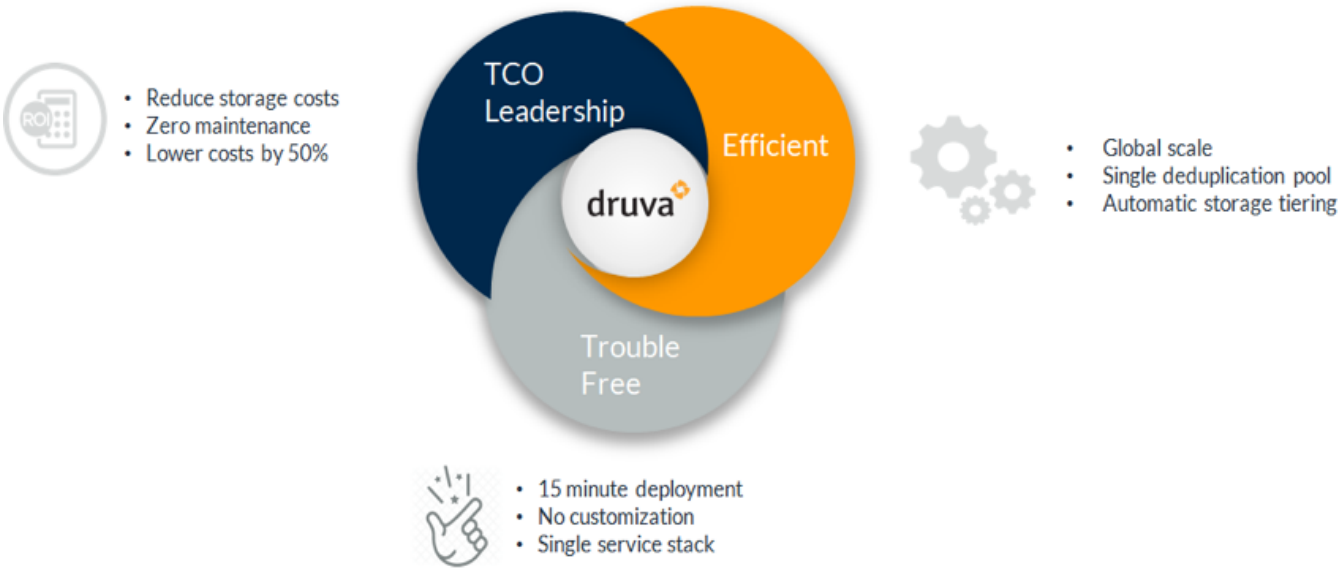
Target Deduplication	Source Deduplication
Pros: • Transparency: point backups to the deduplication appliance and process occurs	Pros: • Speed: first full backup then block-level incrementals after that • Looks directly at files scheduled for backup resulting in higher dedupe ratios • Charge once for dedupe then replicate to many locations • Aggregates the power of multiple sources • Eliminates purchase of hardware appliance(s) • Network efficiency/savings • Faster backups, restores
Cons: • Need to decipher backup container (e.g. NetBackup, CommVault) which are cryptic and difficult to parse • Need to buy multiple dedupe appliances • Development cost to decipher and chunk backup behind proprietary backup format • Full and incremental backups transferred across the network eliminating any network efficiencies	Cons: • Increased computational power required to compute hash, but offset by lower network traffic

Druva

Data is the core to any digital transformation. By simplifying and making data protection more reliable, Druva enables any enterprise to accelerate growth by simplifying data protection and providing data intelligence that accelerates business decisions. Offered on a software-as-a-service platform, Druva offers cloud-based backup and disaster recovery across endpoints, data centers and cloud workloads. It eliminates the need to purchase hardware, software or skilled resources dedicated solely to data protection, reducing total cost of ownership by up to 50%. Built on AWS, customers experience global scale data protection that is secure and offered across all AWS regions.

The figure below provides high level benefits of Druva's Cloud Platform.

Figure 1. Druva Cloud Platform Benefits



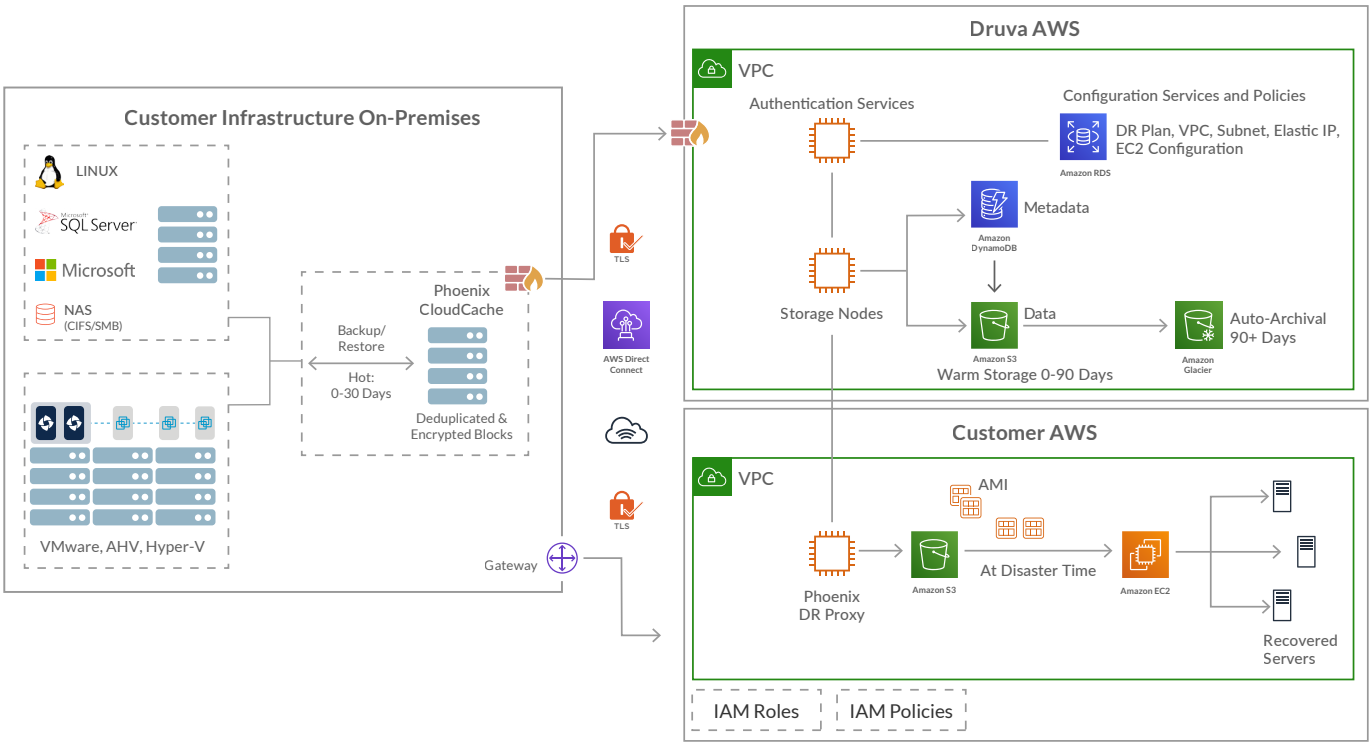
Druva Architecture

Druva is built on AWS and provides customers with access to the latest in high-performance cloud technology, unmatched storage flexibility, data durability and zero-day support for new AWS services. Druva delivers storage efficiencies by leveraging S3, S3IA, Glacier and Snowball Edge.

Druva delivers a single solution that unifies data protection to break up operational silos and simplify administration. For example a unified dashboard across all AWS regions supports data locality requirements while providing IT and business teams with single-click management.

For database services, Druva leverages Amazon's DynamoDB and Relationship Database Service (RDS). For scalability it leverages AWS' load balancer to automatically provision the appropriate amount of EC2 resources for the current backup load. This means that EC2 resources will shift dramatically throughout the day as Druva automatically scales resources up or down. Automatic scaling is completely transparent to the user as they never see or need to manage any backup servers. Customers view their data through a single Druva console.

Figure 2. Druva Architecture



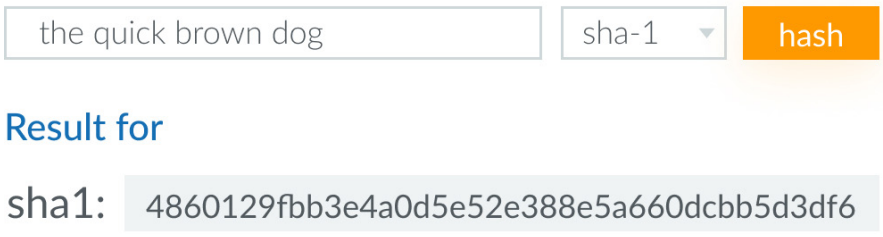
Druva's source side deduplication leverages DynamoDB and RDS. DynamoDB stores the file and deduplication metadata and also performs the deduplication look-ups. DynamoDB also contains the backup and history indexes. RDS stores all the configuration data such as disaster recovery plans, VPCs, subnet, etc.

> Data deduplication has been extremely valuable for Build Group whose data volumes exceed 11 Tbytes. The amount of data transferred over the network is far less with the IT team seeing dramatic reductions in network congestion. Complete backups have gone from a full day to less than half an hour.

How Druva Deduplication Works

Druva globally deduplicates on the client (or source) by chunking files, database dumps, etc into a slice of data that is run through SHA1 to create a hash, shown in Figure 2. A comparison is performed in the hash table to determine if this specific chunk is unique or not unique. If the hash is unique, the data chunk is sent to Amazon's Simple Storage Service (S3) and stored as an individual object. If the hash is not unique, the hash table is updated.

Figure 2. SHA1 example



To minimize the number of chunks, Druva implements incremental forever backups and date stamps. Incremental backup only submits new or modified blocks, files or objects to the chunking process. Druva also uses time stamps to identify files that have been modified.

Storing chunks as objects not only works well with S3 storage, but it also makes capacity maintenance (a.k.a. garbage collection) easier. Using the customer's defined retention policy, Druva identifies objects that are candidates for deletion. A chunk will not be deleted until Druva has verified that it is no longer being referenced by any snapshots.

Chunks that are older than 90 days but still need to be retained per the customer's retention policy, are automatically sent to Amazon Glacier for cost-efficient long term storage.

Druva uses 64K or 1 MB chunk sizes. Hardware-based appliances use 4K-128K chunk sizes. The size of the chunk directly affects the number of records in the deduplication index. Smaller chunk sizes increase the number of records, deduplication time and increases the amount of metadata. To accommodate this growth, customers need to purchase additional appliances which results in fragmentation. and challenges managing multiple deduplication indexes.

The Port of New Orleans has benefited from more efficient backup workflows and restore times. Previously taking between 4 to 8 hours per system, backups and restores now take 30 minutes or less.

Summary

Data deduplication is one of the technologies changing the economics of data protection and management and has quickly become a critical component for any organization's data protection strategy. However, just how much efficiency is achieved depends on the type of deduplication implemented.

Druva uses source global deduplication to deliver compelling value to business:

- Source deduplication delivers network efficiencies for faster backups and restores
- Single dedupe index ensures one copy of data is stored—across all global backups
- Faster backups and restores with less data being transmitted across the network
- Eliminates management challenges and manual decisions to migrate data when organizations outgrows hardware
- Leveraging AWS micro services and storage provides unmatched data durability, storage flexibility and zero-day support for new AWS services

Ready to take it for a spin? Download a trial and experience the benefits of Druva's data protection for the cloud era.

druva

Sales: +1 800-375-0160 | sales@druva.com

Americas: +1 888-248-4976
Europe: +44 (0) 20-3750-9440
India: +91 (0) 20 6726-3300
Japan: +81-3-6890-8667
Singapore: +65 3158-4985
Australia: +61 1300-312-729

Druva™ delivers data protection and management for the cloud era. Druva Cloud Platform is built on AWS and offered as-a-Service; customers drive down costs by up to 50 percent by freeing themselves from the burden of unnecessary hardware, capacity planning, and software management. Druva is trusted worldwide by over 4,000 companies at the forefront of embracing cloud. Druva is a privately held company headquartered in Sunnyvale, California and is funded by Sequoia Capital, Tenaya Capital, Riverwood Capital, Viking Global Investors and Nexus Partners. Visit Druva and follow us @ druvainc.