



data deduplication in the enterprise

the challenge

Storage requirements for organizations are growing exponentially, with digital information being created or replicated at 60% or greater compound annual growth rates. That's a 10 fold increase in the amount of data a company needs to manage over a five year period. How do you keep pace with that kind of growth when power, cooling and floor space simply can't grow at the same rate? How do you manage all that new data while maintaining IT Opex and Capex budgets?

so what's the answer?

How do you keep pace with the phenomenal growth of data in the environment? By only storing unique data once – and treating data as byte patterns or blocks rather than files or objects. While applications may own data, how that data is stored can be virtualized so that applications aren't impacted, while dramatically reducing the amount of real physical storage required. This virtualization replaces duplicate byte patterns or blocks with pointers (metadata), without proprietary API's or modifications to applications; and this naturally reduces costs and management overhead.

what is it?

Deduplication is the process by which duplicate data is removed from a data stream or repository, and replaced by reference pointers.

- Data is segmented into objects, files, chunks, blocks, sub-blocks, or variable sized blocks
- An algorithm searches these segments looking for duplicate data
- When a duplicate segment is detected, it is released (freed) and its reference pointer modified – thus any given unique data segment will have multiple reference pointers

how is it implemented?

Depending on the specific challenges within your environment, implementing a deduplication strategy may take different, and in many cases multiple, forms.

deduplication at source

This is achieved by analyzing data at the source – an application server, file system, etc – and either deduplicating it on the primary storage volume (production file-systems), or deduplicating content for the purposes of transmission and/or storage on another device. Various databases are used to track unique data elements prior to transmission. This can reduce local primary storage requirements, network bandwidth for data transmission, and remote offsite storage capacity for backup/recovery and disaster recovery. Generally speaking, block based deduplication is used for primary production systems and is performed as a scheduled task. Sub-block or byte-pattern deduplication is more CPU intensive and thusly is typically reserved for offsite backup and recovery solutions.

inline deduplication engines

This is a more intense operation that needs powerful engines to provide fast deduplication performance while data is being stored, retrieved, and transmitted. One big advantage of this technology is deduplication occurs in real time as data is read or written – which means things like replication can also be done on stored deduplicated data in real time. No additional storage is required as “temporary scratch” space, which can significantly increase the number of spindles and capacity required for optimal performance. This technology lends itself well to both target storage devices for backup and recovery, with replication for offsite copies as well as in-band network deduplication devices that reduce network bandwidth requirements, dramatically improving application performance and enabling data consolidation.

post deduplication engines

When implemented as a post-process method, performance is not compromised. Deduplication occurs after the data is stored, so that it doesn't interfere with applications or users accessing the content, or slow down storage of data on the device. Post-process implementations are typical for Virtual Tape Library appliances and require a staging area for data to be written to first. This is also the ideal methodology for doing data deduplication of primary production filesystems and LUNs, since the production application wouldn't be impacted.

where can this benefit me?

Significant savings in storage capacity and network bandwidth can be achieved through the implementation of this technology at a variety of touch points within an organizations IT infrastructure. These include:

- Backup Applications
- Replication Optimization
- Disaster Recovery & Tape Elimination
- Application Protection Utilities – i.e, Oracle RMAN, DB2, SQL Server, SAP, Exchange
- Virtualization – OS Images, Desktop (VDI)
- ILM – Project Archives, Engineering Versions, Compliance Retention, Home Directory
- Archiving Applications – CommVault, Symantec, Hitachi Data Systems, EMC, Arkivio, Atempo,
- Tiered Storage Solutions – F5 Acopia, Brocade StorageX, Archiving Solutions
- WAN Acceleration and Data Reduction

how can we help?

As a leading provider of technology solutions and consulting services, Scalar Decisions Inc. is well versed in a variety of approaches and solutions from multiple vendors for leveraging the significant benefits that data de-duplication can bring to Information Technology. Through further dialog, workshops, and technology reviews, Scalar Decisions Inc. will work with you to understand your specific requirements, select the best solution and technology to meet your needs, and deliver compelling ROI to the business.

For more information, please contact sales@scalar.ca or visit us at www.scalar.ca

