

Industry Trends and Technology Perspective White Paper

Business Benefits of Data Footprint Reduction

Why and how reducing your data footprint provides a positive benefit to your business and application service objectives

By Greg Schulz

Founder and Senior Analyst, the StorageIO Group

July 15th, 2007

The information necessary to support your business in timely and effective decision making and in maintaining a competitive advantage has an impact on your data footprint. This paper looks at various techniques, and, in particular, how data compaction compression technologies can be applied to various types of data and IT functions to help reduce your data foot print to improve energy consumption and, enhance existing storage resource utilization while addressing other IT Infrastructure Resource Management (IRM) inefficiencies as well as enhancing overall application service levels.

Introduction

Organizations of all sizes are generating and depending on larger amounts of data that must be readily available and easily accessible. This growth in data results in an ever increasing data footprint; that is, more data is being generated, copied, and stored for longer periods of time. Consequently, IT organizations have to effectively manage more infrastructure resources, including servers, networks and storage, to insure that data is protected in a timely manner while at the same time providing adequate performance and capacity and securing data for access when needed.

Background

Your data footprint is the total data storage needed to support your various business application and information needs. Your data footprint may, in fact, be larger than how much actual data storage you have, or, as in the following example, you may have more aggregated data storage capacity than actual data.

As an example, you have 2 TBytes of Oracle database instances and associated data, 1 TByte of Microsoft SQL data, 2 TByte of Microsoft Exchange Email data, 4 TByte of general purpose shared NFS and CIFS Windows based file sharing resulting in 9 TByte (2+1+2+4) of data, however, your actual data footprint could be much larger. The 9 TB simply represents the known data or how storage is allocated to different applications and functions. If the databases are sparsely populated at 50%, for example, only 1 TByte of Oracle data actually exists while occupying 2 TByte of storage capacity.

Determining your data footprint

- A general approach is to simply add up all of your on-line, near-line and off-line data storage (disk and tape) capacity.
- Another approach is to add up the storage capacity space required by different applications plus space needs for operating system and application software, temporary and work space plus requirements for disk mirroring, RAID, backup and snapshots as well as growth and free space.
- Exercise caution to make sure you are not double counting, yet include overhead or infrastructure needs for backup, data conversions, import / export and other IRM maintenance functions.

Assuming, for now, that in the above example the capacity sizes mentioned are fairly accurate to the actual data size based on how much data is being backed up during a full backup, your data footprint would include the 9 TByte of data as well as the on-line (primary), near-line (secondary), and off-line (tertiary) data storage configured to your specific data protection and availability service requirements. For example, if you are using RAID 1 mirroring for data availability and accessibility, in addition to replacing your data asynchronously to a second site where the data is protected on a RAID 5 based volume with write cache, as well as a weekly full backup your data footprint would then be at least $(9 \times 2 \sim \text{RAID 1}) + (9+1 \sim \text{RAID 5}) + (9 \sim \text{Full backup}) = 37 \text{ TByte}$.

Your data footprint could be even higher than 37 TByte in this example if we also assume that daily incremental or periodic snapshots are performed throughout the day in addition to extra storage to support application software, temporary work space, operating system files including page and swap, not to mention room for growth and whatever free space buffer used for your environment.

As can be seen from this example, 9 TBytes of actual or assumed data can rapidly expand into a larger data footprint that only compounds as your applications grow to support new and changing business needs or requirements. Note that the above scenario is rather simplistic and does not factor in how many copies of duplicate data may be being made or backup retention, size of snapshots, free space requirements and other items that contribute to the expansion of your data footprint.

Reducing your data footprint

While storage capacity has, in fact, become less expensive, as your data footprint expands, more storage capacity and storage management, including software tools and IT staff time, are required to care for and protect your business information. By more effectively managing your data footprint across different applications and tiers of storage, you can enhance application service delivery and responsiveness as well as facilitate more timely data protection to meet compliance and business objectives.

Reducing your data footprint can help reduce costs or defer upgrades to expand server, storage and network capacity along with associated software license and maintenance fees. Maximizing what you already have using data footprint reduction techniques can extend the effectiveness and capabilities of your existing IT resources, including power, cooling, storage capacity, network bandwidth, replication, backup, and archiving and software license resources.

From a network perspective, by reducing your data footprint or its impact, you can also positively impact your SAN, LAN, MAN and WAN bandwidth for data replication and remote backup or data access, as well as move more data with existing available bandwidth. Additional benefits of maximizing the usage of your existing IT resources include:

- Deferring hardware and software upgrades, maximizing usage of existing resources
- Enabling free space to facilitate consolidation and data migration to energy effective technologies
- Shortening time required for data protection, including file system scans and data movement
- Reducing your power and cooling requirements by increasing utilization of existing storage
- Expediting data recovery and application restart for DR scenarios
- Environmental improvement to maximize your energy power consumption and cooling
- Lowering the impact from file system scans for backup and other overhead functions
- Reducing the impact of IRM related overhead functions, including file system scans for backup
- Less exposure during RAID rebuilds due to faster copy times due to denser data

	Archiving	Compression	De-Duplication
Known As	Data pruning or copy and deletion of old data	Compaction	Single instance storage (SIS), commonality factoring, data differencing
When to use	Structured (Database), Email and Un-structured	On-line (Database, Email, Files), Backup or Archive	Backup or Archiving
Characteristics	Software to identify and remove un-used data from completed projects or for compliance	Reduce amount of data to be moved (transmitted) or stored on disk or tape.	Eliminate duplicate files or file content observed over a period of time to reduce data footprint
Examples	Database, Email, Un-structured file solutions with archive storage	Host software, disk or tape, (network routers) and compression appliances	Backup and archive target devices and VTLs, specialized appliances
Caveats	Time and knowledge to know what and when to archive and delete, data and application aware	Software based solutions require host CPU cycles which negatively impacts application performance	Works well in background mode for backup data to avoid performance impact during data ingestion

Figure-1: Data footprint reduction approaches and techniques

Figure-1 shows techniques and approaches that can be combined in various ways to complement each other to address and reduce your data footprint. IT organizations have taken different approaches to addressing the challenges associated with a growing data footprint while balancing service delivery (performance, availability, capacity, compliance) with cost, including available management man power, operating expense (opex) along with capital expense (capex), while insuring that compliance and business continuance (BC) or disaster recovery (DR) requirements are being met. While DR and compliance have been in the news recently, along with data security, another topic that is gaining attention is green data storage and green IT infrastructure in general and, specifically, reducing electrical power and cooling consumption or doing more with less.

For some organizations the solution to reducing data footprint involves restricting the use of storage. This basically limits users and applications in how much data can be retained or made available. Some examples range from limiting database size to placing restrictions on email box size and user disk space quotas. While limits and quotas can have their place, the implementation of these quotas and limits should not impede or hinder the ability of an organization to compete in the market place. Another approach is to simply add more hardware. After all, disk storage capacity continues to grow while prices drop. Keep in mind that while the disk hardware and associated controllers can be relatively cheap and even energy efficient to operate, they still require software and management, including backup and other functions, which result in personnel and software costs.

Archiving of un-used data

If you do more than simply move the data to a different location, data archiving can have one of the greatest impacts on reducing your data footprint, for storage in general but particularly for on-line and primary storage. For example, if you can identify in a timely manner what data can be removed after a project is completed or what data can be purged from a primary database or older data migrated out of active email databases, you should realize a net improvement in application performance as well as available storage capacity.

A challenge with archiving is having the time and tools available to identify what data should be archived and what data can be securely destroyed when no longer needed. Further complicating archiving is that knowledge of the data value is also needed; this may well include legal issues as to who is responsible for making decisions on what data to keep or discard. If you can invest in the time and software tools, as well as identify which data to archive to support an effective archive strategy, the returns can be very positive towards reducing your data footprint without limiting the amount of information available to your business.

What about ILM?

Information Lifecycle Management (ILM) is a term widely used and misunderstood to mean a variety of different things. The spirit and intent of ILM as a technique or paradigm, not as a product, is to more effectively manage and enable data footprint reduction along with effective data and storage management. A challenge with ILM is that it means different things from archiving, tiered disk drives or automatic policy based data all of which require extensive software and personnel time to effectively implement.

Single instance storage (SIS) and data De-duplication

Single instance storage, also known as data de-duplication, assumes that duplicate files exist on a server or storage system being backed up, and that over time the same un-changed files get repeatedly backed up. SIS work by normalizing the data being backed up and subsequently stored- that is, instead of storing each file containing the same data, keep one copy of the actual data and maintain multiple pointers to the data representing the various files being backed up.

SIS is similar to Point-in-Time (PIT) copies. These are pointer-based snapshots with pointers maintained to changed data instead of copying entire files. The benefits of PIT pointer based snapshots are speed of data protection and less storage required for rapid retrieval of data. SIS approaches trade processing time to ingest and eliminate duplicate data for a savings on storage capacity to store backed-up data. This assumes that there is a high degree of commonality and repeating data files being backed up. Consequently, SIS and data de-duplication solutions perform best when deployed for and in support of backup operations and to a lesser degree for archiving. These solutions are not practical for on-line data applications today. Some SIS and data de-duplication enabled solutions, such as virtual tape libraries, also combine basic data compression to further reduce data footprint requirements across a broader scope and range of applications and data formats.

Data compression

Compression is a proven technology providing immediate and transparent relief to move or store more data effectively not only for backup and archive but also for primary storage. Data compression is widely used in IT as well as in consumer electronic environments. It is implemented in hardware and software to reduce the size of data to create a corresponding reduction in network bandwidth or storage capacity. For example, if you have ever zipped a file, listened to an MP3, or watched HDTV you are relying on compression technologies. Compression technology is very complementary to archive, backup and other functions, including supporting on-line, primary storage and data applications. For example, compression is commonly implemented in several locations including databases, email, operating systems, tape drives, network routers and compression appliances to help reduce your data footprint.

Some data de-duplication solutions boast spectacular ratios for data reduction given specific scenarios such as backup of repetitive files while providing little value over a broader range of applications. This contrasts with traditional data compression approaches that provide lower, yet more predictable and consistent data reduction ratios over more types of data and application, including on-line and primary storage scenarios. For example, in environments where there is little to no common or repetitive data files, data de-duplication will have little to no impact while data compression generally will yield some amount of data footprint reduction across almost all types of data. Some data de-duplication solution providers have either already added, or have announced plans to add, compression techniques to compliment and increase the data footprint effectiveness of their solutions across a broader range of applications and storage scenarios, attesting to the value and importance of data compression to reduce your data footprint which can not be ignored.

What to look for in a data footprint reduction solution

Many vendors' sales pitches lead with messages focused on reducing opex or capex costs or doing more with less, which for many environments is a good thing. Doing more with what you already have can be interpreted as "doing more with less," however, it also has a different meaning; for example, increasing the storage capacity utilization of existing disk and tape based storage systems, leveraging virtualization techniques to consolidate server workloads, maximizing available power and cooling capacity with either more efficient technology, or, to make room for new technology, maximize the storage capacity utilization of existing devices while not consuming more network or host server CPU processing cycles.

There are many different attributes to consider when evaluating data footprint reduction technologies. Which features are the most important for you will depend on your environment, needs and requirements. Figure-2 shows desirable attributes and functionalities to factor into your evaluation of reduction techniques and, specifically, data compression solutions.

Functionality	Archive	Compression	De-duplication
Data footprint reduction effectiveness	Effective for reducing data footprint however requires more effort and cost to achieve given results	Effective across backup, archive as well as on-line and primary applications and all tiers of storage	Good for backup of same or similar data and files, not as effective for dissimilar data or on-line primary storage
Performance	Performance impact on servers, storage and network to copy and delete archive data.	Appliance based solutions eliminate performance penalty to host servers, network or storage system	May introduce performance bottleneck during in-line processing of backups or during bulk data recovery
Ease of use	Requires installation and integration of application aware software and archive policies	Works with existing applications and technologies to reduce learning curve	May require learning curve of new technology as well what scenarios work best with the technology
Transparency	Requires application specific and aware archive software along with usage policy rules and management	Transparent across existing applications, maximizes your existing technology investment including across all storage tiers	Introduces a new layer of technology that may or may not co-exist with your existing storage and management tools.
Flexibility	Eliminates large amount of data, however complicated and involved from a time perspective to implement	Transparency across existing applications and all tiers of storage	May impose performance delay during bulk data restoration while waiting for data to be expanded
Management	Introduces additional software and or archive storage medium to manage along with policy definition and management	Look for solutions with ability to select which files or shares to compress along with monitoring and analysis tools	May introduce a new tier of storage or storage software to manage along with integration with backup or archive off-line devices
Scalability	Number of clients, databases, tables, records, files or mailboxes among other objects? Look at the effective data ingestion rate for your class of data	Look for solutions that do not introduce instability in terms of performance bottlenecks during everyday processing as well as during BC or DR events	Number of concurrent backup streams that can be processed concurrently without disruption to backup or archive jobs
High Availability	Archiving by itself does not have a direct impact on availability other than generally reducing the amount of data to be protected	Look for solutions that support transparent failover to redundant appliances for continued access of data along with optional software for DR purposes	Avoid performance bottlenecks due to single points of failure as well as delays in restoring multiple data files or storage volumes concurrently

Figure-2: What to look for in a data footprint reduction solution

Recommendations to reduce your data footprint

In the storage marketplace, there has been a vertical or product-centric focus on how to reduce data footprints including, as an example reducing the amount of stored data and associated storage capacity to improve backup objectives. Another focus has been on promoting fixed content archiving, ediscovery content search and data indexing for legal, regulatory or other compliance purposes. There are many opportunities to reduce your data footprint to improve overall delivery of service and enhance

management and the ability to positively impact spending on hardware and software, not to mention recurring software maintenance fees for storage tools.

An issue to consider is how much delay or resource consumption can you afford to use or lose to achieve a given level of data footprint reduction? For example, as you move from coarse (traditional compression) to granular, such as data de-duplication or single-instancing, more intelligence, processing power, or off-line post processing techniques are needed to look at larger patterns of data to eliminate duplication.

Deploy a comprehensive data footprint reduction strategy combining various techniques and technologies to address point solution needs as well as your overall environment, including on-line, near-line for backup, and off-line for archive data. Following are some general recommendations and suggestions to help address your growing data footprint, some of which should be common sense; all will vary by the size and scope of your particular environment, application and service requirements.

- ✓ If you are concerned enough to be evaluating other forms of data footprint reduction technologies for future use, including archiving with data discovery (indexing, ediscovery) or data de-duplication techniques, leverage appliance-based compression technology for immediate relief to maximize the effectiveness and capacity of existing storage resources for on-line, backup, and archive while complimenting other data footprint reduction capabilities.
- ✓ Understand how other IRM functions, including backup, archive, DR/BC, virus scans, encryption and ediscovery along with indexing for search, interact with data footprint reduction technologies.
- ✓ Maximize the usage of your existing IT infrastructure resources without introducing complexity and costs associated with added management and interoperability woes. Look for solutions that compliment your environment and are transparent across different tiers of storage, business applications and IRM functions (backup, archive, replication, on-line).
- ✓ Data archiving should be an ongoing process that is integrated into your business and IT resource management functions as opposed to being an intermittent event to free up IT resources.
- ✓ Get a handle on your data footprint and its impact on your environment using analysis tools and/or assessment services. Develop a holistic approach to managing your growing data footprint- look beyond storage hardware costs, factor in software license and maintenance costs, power, cooling and IT staff management time.
- ✓ Leverage data compression as part of an overall data footprint reduction strategy to optimize and leverage your investment in your existing storage across all types of applications.

Conclusion

There are several different techniques that can be used individually to address specific data footprint reduction issues or in various combinations to implement a more cohesive and effective data footprint reduction strategy. The benefit of a broader, more holistic, data footprint reduction strategy is to address your overall environment, including all applications that generate and use data as well as IRM or overhead functions that compound and impact your data footprint.

Reducing your data footprint has many benefits, including reducing or maximizing the usage of your IT infrastructure resources such as power and cooling, storage capacity, network bandwidth while enhancing application service delivery in the form of timely backup, BC/DR, performance and availability. If you do

not already have a data footprint reduction strategy, now is the time to develop and begin implementing one across your environment.

Look to combine multiple technologies and techniques to address your various data footprint challenges to maximize your IT resources while reducing management costs and complexity. Data compression using appliances that are transparent to applications and across different tiers of storage is a timely and effective means to achieve immediate relief and complement your existing technologies and investment,

About the author

Greg Schulz is founder and senior analyst of the StorageIO Group and author of the book *Resilient Storage Networks — Designing Flexible Scalable Data Infrastructures* (Elsevier Digital Press).

All trademarks are the property of their respective companies and owners. The StorageIO Group makes no expressed or implied warranties in this document relating to the use or operation of the products and techniques described herein. The StorageIO Group in no event shall be liable for any indirect, inconsequential, special, incidental or other damages arising out of or associated with any aspect of this document, its use, reliance upon the information, recommendations, or inadvertent errors contained herein. Information, opinions and recommendations made by the StorageIO Group are based upon public information believed to be accurate, reliable, and subject to change.