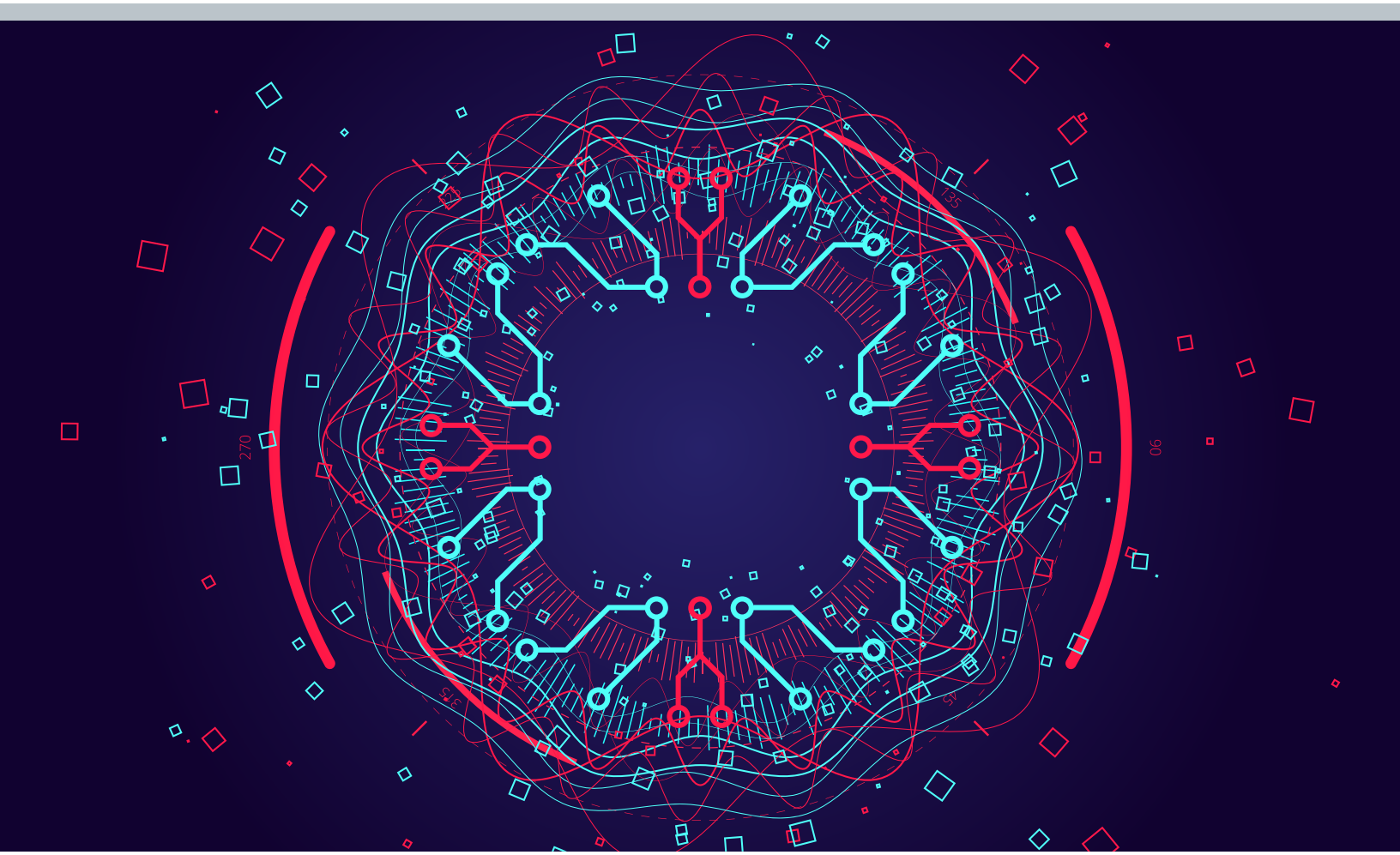
PAPER

Redmond MAGAZINE



DATA GROWTH: DATA RECOVERY'S WORST NIGHTMARE

By Nick Cavalancia

EXAGRID®

VEEAM

www.exagrid.com

It's simple math: the larger the data set, the longer it takes to back up or recover data. Plus, more data means more storage, and that can be costly. It's an unfortunate truth for IT, as you're tasked with ensuring application and system availability while grappling with continually growing data. It's also assumed that you'll simultaneously address increasing retention requirements and reduce recovery times—all while working within a limited budget. That's a tall order for any IT pro.

None of us have ever heard of an IT department that is shrinking the number of supported applications, systems, and the size of the resultant data set. For every one of us, the cold harsh reality is that more is being added onto IT's plate day-by-day, quarter-by-quarter.

All this becomes even more of a challenge when it comes to data recovery. As the number of systems, applications, and amount of data increases, you're left with the arduous task of prioritizing each, determining the best way to properly protect and recover them (if that's even possible…), while also meeting the SLA and application availability requirements of your organization.

All this comes at an expense—certainly budgetary costs are a constraint, but the organization's ability to recover, IT's ability to deliver, and even your sanity may all be impacted.

What would be immeasurably helpful is an actual reduction in the amount of data to be backed up, right? If it were possible to back up less data, it would be easier to plan backup windows and recovery times, meet recovery objectives, and maximize business continuity. To accomplish this, you need to either reduce production systems and data, or somehow reduce the amount of data for which you're responsible. Both seem like nothing short of impossible.

So, how can you control data sprawl while actually improving your ability to back up and recover?

The answer lies in reducing the size of your backups. While this sounds simple, actually accomplishing this is not. In this paper, we'll take a look at the problem of data growth, its impact on backup storage, and the hidden costs associated with the growth of data. We'll also look at the use of deduplication as part of your backup and recovery efforts, and

how you can leverage it to ensure the smallest backup and data recovery sets possible.

Where's the greatest source of increased storage in backups?

IT is experiencing a surge in the amount of data it is responsible for. You're experiencing it in the form of application data, the sheer number of applications and systems, and the requirement for tighter recovery objectives (which result in more frequent backups). But, it's not a sudden development that your organization has exponentially more customers in the CRM database, is sending more emails with vastly larger attachments, etc. It's something more fundamental to how businesses are operating today.

WHAT WOULD BE IMMEASURABLY HELPFUL IS AN ACTUAL REDUCTION IN THE AMOUNT OF DATA TO BE BACKED UP, RIGHT?

Overall, the world is seeing data double every two years[1]. For organizations like yours, the surprising reason you're seeing growth isn't business data—the day-to-day Microsoft Word and Excel documents, emails, etc. In actuality, it's the growth in data used for analytics and decision making. Sensor data from IoT devices alone is expected to grow by 50 times during the next few years[1]. Human data (think user activity log data) is already mainstream in IT and is expected to grow in a similar fashion to that of sensor data. Business data continues to grow rapidly as well but nowhere near the rates of the other data sources mentioned.

Think about it: IT needs to deliver applications that give business owners actionable insight. For example, Sales, Marketing, Production, Customer Service, and IT all expect to have solutions in place that turn information into intelligence, and intelligence into insight. Dashboards, reports, alerting, etc. all rely on very large data sets—in fact, the larger the better in many cases. Just take the use of a Security Event Information Management (SEIM) solution. If you have one, it makes the case that even IT needs lots of data from disparate sources in one place to gain insight into where there are security issues. Add to this the increasing use of artificial intelligence (AI) technologies as part of newer solutions, and the reliance on having historical data to feed AI increases.

[1] InsideBigData, The Intelligent Use of Big Data on an Industrial Scale (2017)

So, how does all this impact backup and data recovery?

More Data, Larger Backups, and Data Recovery

As organizations increase their reliance on data, IT is faced with increasingly larger backup data sets. As this trend shows no sign of changing anytime soon, there are several ways this impacts you—some obvious and some not so obvious:

- **Backup and Recovery Windows** – If you're doing backup correctly, you've started with establishing recovery objectives, and creating policies for both backup and recovery. If you have a 15-minute recovery time objective for a given data set, it becomes increasingly difficult to meet that SLA as the data set grows.

- **Recoverability** – If, at some point, a data set grows to an unmanageable size, the question of whether it's possible to recover it in the time allotted—or at all—is raised. In some instances, the method of data protection needs to change (e.g., moving from backup sets to actual copies of virtual machine data sets and applications to enable near instantaneous recovery.)

- **Retrieval Times** – Backups of larger data sets often reside on secondary storage media since they don't change much and are assumed to be recovered infrequently. When on secondary storage, the data isn't necessarily accessible for instantaneous recovery unless it's designed to quickly recover the more recent backup copies like ExaGrid. In a standard secondarily scenarios like Data Domain, additional time is necessary to retrieve the data for recovery.

- **Retention** – As your data grows, depending on the data type and your organization's industry, retention requirements may dictate that the exponentially growing data set be retained for years, increasing your backup and storage requirements beyond original intentions or plans.

IF YOU'RE DOING BACKUP CORRECTLY, YOU'VE STARTED WITH ESTABLISHING RECOVERY OBJECTIVES, AND CREATING POLICIES FOR BOTH BACKUP AND RECOVERY.

In addition (and it's been hinted at already in the bullets above), the inclusion of massive data sets changes the conversation about your entire data recovery strategy. The current data recovery strategy around recovering tier 1 workloads and their data may need to change in order to meet both long-term data retention for compliance reasons, and the need to ensure a quick recovery to meet application availability SLAs.

There are other costs beyond those you're already experiencing with data growth.

THE HIDDEN COST OF DATA GROWTH: WHAT'S IT REALLY COSTING YOU?

There's a logical assumption that as your data grows, there will be a tangible rise in support costs. Storage hardware is the obvious first source of increased cost—the addition of disk drives and arrays are evident when planning for storage growth. Certainly, the retiring of spinning disks and moving to solid state drives that have a smaller footprint and larger capacity may offset some costs. But even so, as data grows, so will hardware costs—even when using state-of-the-art storage.

There are other costs attributable to the cost of data growth that you may not have considered. Some are similar to those previously mentioned while others are repercussions of data growth:

- **Operational Overhead** – Regardless of the hardware you have in place, someone needs to manage your storage environment. As that environment becomes more complex due to added arrays, new drives, the use of tiered storage, etc., the work necessary to keep that environment highly available increases.

- **Retention and Compliance Overhead** – If the growing data is subject to retention requirements, the cost of retaining that data over time will equally increase. The use of tiered storage should help to offset some of the added cost, but you should expect that more retention equates to more storage-related costs.

- **Recovery Time (RTO)** – It may be necessary to retrieve the entirety of a very large data set for a critical application to function. Depending on the performance of the hardware and software in use, and the ability to quickly rehydrate your data, you may be looking at a lengthy restore time, which can be quite costly from an operations and productivity perspective.

As your data grows, it requires a larger storage footprint, and this increases the cost of not just storing data, but also managing, maintaining, and recovering it.

So, how can you minimize the true amount of data being backed up, stored, and recovered?

DECREASING STORAGE WITH DATA DEDUPLICATION

The answer lies with deduplication: a process that eliminates matching copies of redundant data within a data set. While simple in concept, the practical application is quite complex—the optimal algorithms used for one type of data, such as a VM database, are not necessarily optimal for a completely different data type. Deduplication is complex and the approaches are varied. For deduplication to have maximum impact, it needs to be expertly architected

AS YOUR DATA GROWS, IT REQUIRES A LARGER STORAGE FOOTPRINT, AND THIS INCREASES THE COST OF NOT JUST STORING DATA, BUT ALSO MANAGING, MAINTAINING, AND RECOVERING IT.

and executed; for instance, it's actually possible to have both software- and hardware-based deduplication in play simultaneously and end up with a larger data set than you started with!

When properly implemented, data deduplication balances a number of factors:

- **Data Size** – The initial size of the object being processed, as well as whether its size changes materially between backups, can alter an algorithm's ability to deduplicate quickly and effectively.

- **Backup Size** – You'd think the primary goal of deduplication is to achieve the smallest data size possible. While a small backup size is important, keep in mind that you may sacrifice time not only when reducing the data size but also when rehydrating the data during recovery.

- **Performance** – This, in essence, is the flip side of backup size. It's necessary to maintain an equilibrium between small backups and fast backup and recovery. Too small and your performance will suffer. Too much focus on performance and your backup sizes will be larger than desired. Superior deduplication solutions effectively leverage these factors for optimal performance.

Of these three factors, the most important is likely performance. When performing a recovery, no one cares about how small the data set is; the only concern is how quickly the recovery can occur. As part of the recovery process, most vendors offering deduplication must first rehydrate the data when the recovery is requested, which makes recoveries take longer to complete. Some vendors architect their systems so that the last backup created is not only stored as deduplicated data, but the non-deduplicated data is also stored in its native form. This way, the most recent backup (which usually holds the most requested data in any organization) can be quickly recovered because the need for time-consuming data rehydration is eliminated.

In the case of backups, there are generally two ways data can be deduplicated: via backup software or via storage hardware. The benefit of using backup software is the intelligent selection of the best deduplication algorithm based on the data type being backed up. The drawback of using software is that it's usually not as fast as hardware-based deduplication. In general, any time you

can leave deduplication to hardware, you're better off, as software-based deduplication is very processor and memory intensive. But, some of you may not have a storage solution that provides deduplication.

WAKING UP FROM THE DATA RECOVERY NIGHTMARE WITH DEDUPLICATION

The amount of data you're responsible for will only grow over time, which means that your DR efforts will become more challenging as time goes on. Optimizing your application SLAs based upon your data, criticality, and retention requirements while leveraging the deduplication provided by both your backup solution and storage hardware will help improve your capabilities.

Deduplication can solve a good percentage of your secondary storage capacity issues—both from the perspective of backup size as well as storage and on-site retention for recovery. It is important to consider that simply leveraging deduplication won't help you achieve optimal results. Understanding how each vendor balances performance, reduces the amount of data stored, and how different vendors complement your backup infrastructure will help you make the most of your infrastructure today while preparing you for the challenges of managing data in the future.

Find out more

https://www.exagrid.com

Nick Cavalancia is founder & chief techvangelist at Techvangelism. Nick has more than 20 years of enterprise IT experience, and is an accomplished consultant, speaker, trainer, writer, and columnist. He has several certifications including MCSE, MCT, Master CNE and Master CNI. He has authored, co-authored and contributed to over a dozen books.