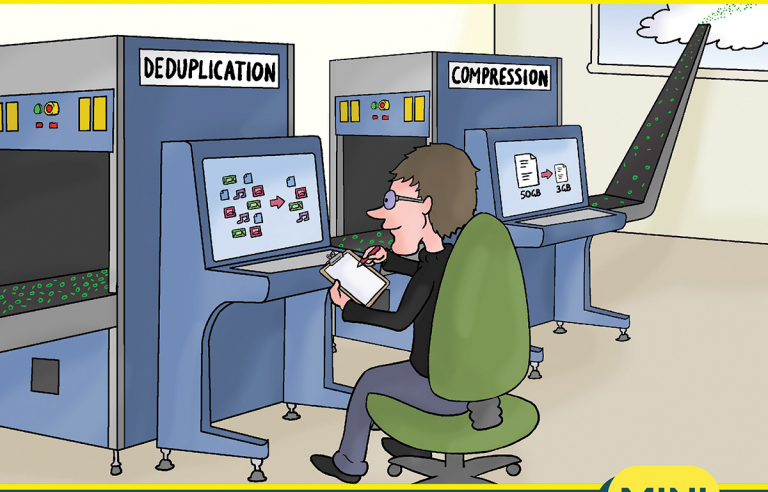
ConversationalGeek

Conversational Backup Optimization for the Cloud

Nick Cavalancia (4-time Microsoft Cloud and Datacenter MVP)



MINI Edition

Learn about:

- Why backing up data to the cloud is a continually-growing challenge
- Ways to optimize your backup data to speed up backups and lower costs

Sponsored by Quest™

Quest creates software solutions that make the benefits of new technology real in an increasingly complex IT landscape. From database and systems management, to Active Directory and Office 365 management, and cyber security resilience, Quest helps customers solve their next IT challenge now. Around the globe, more than 130,000 companies and 95% of the Fortune 500 count on Quest to deliver proactive management and monitoring for the next enterprise initiative, find the next solution for complex Microsoft challenges and stay ahead of the next threat. Quest Software.

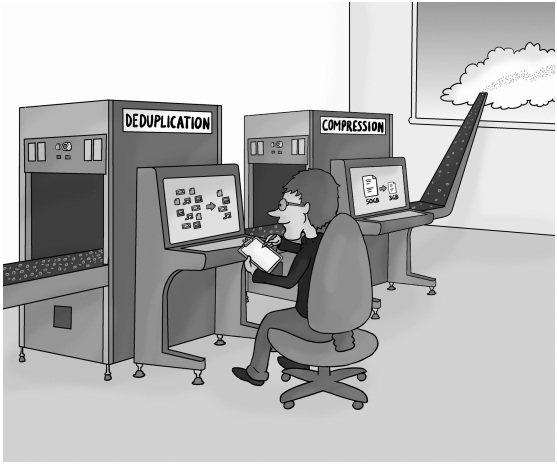
Where next meets now.

Quest

For more information, visit www.quest.com.

Conversational Backup Optimization for the Cloud (Mini Edition)

by Nick Cavalancia



Conversational Backup Optimization for the Cloud (Mini Edition)

Published by Conversational Geek® Inc.

www.ConversationalGeek.com

Trademarks

Warning and Disclaimer

Additional Information

For general information on our other products and services, or how to create a custom Conversational Geek book for your business or organization, please visit our website at www.ConversationalGeek.com.

Publisher Acknowledgments

All of the folks responsible for the creation of this book:

Author:	Nick Cavalancia
Project/Copy Editor:	Pete Roythorne
Content Reviewer(s):	Bennett Klein

The “Conversational” Method

We have two objectives when we create a “Conversational” book. First, to make sure it’s written in a conversational tone so that it’s fun and easy to read. Second, to make sure you, the reader, can immediately take what you read and include it into your own conversations (personal or business-focused) with confidence.

“Geek in the Mirror” Boxes

We infuse humor and insight into our books through both cartoons and light banter from the author. When you see one of these boxes it’s the author stepping outside the dialog to speak directly to you. It might be an anecdote; it might be a personal experience.

Within these boxes I can share just about anything on the subject at hand. Read ’em!

Getting the Most Out of Cloud Backups

"I need you to back all that up into this please."

You know all that data represented by your operational workloads, servers, critical endpoints, data in use, data in archive, DevOps environments, etc.?

Yeah, you're responsible for backing all that up.

The bad news is that the sheer amount of data your organization creates (and therefore becomes your problem) is only going to grow.

Today, it's estimated that the data created in 2022 is somewhere around 100 zettabytes (that's one hundred billion terabytes!) in total[1]. It's estimated that by 2025, over 100 zettabytes of data will be stored in the cloud, with another 100 zettabytes stored everywhere else (including on-prem infrastructure, endpoints, IoT devices, and more)[2]. This means both your operational data, and their related backups are going to grow.

And where will most of this data reside?

Anecdotally, you can guess it's in the cloud, but there's real data to support that assumption.

---

[1] Statista, "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025"

[2] Cybersecurity Ventures, "The World Will Store 200 Zettabytes Of Data By 2025"

In 2015, only 30% of corporate data was stored in the cloud – in 2022 that number has doubled to 60%[3] with consistent growth each year.

The result is the assumption that your organization's use of the cloud for operations and, therefore, your use of cloud storage as part of your backup and disaster recovery efforts is also going to grow.

Backing Up All Your Data is a Challenge

In essence, you're at the mercy of the growing amount of backup data you must retain and manage, which determines a lot of things these days:

- How much storage you consume
- How much you pay for storage

[3] Statista, "Share of corporate data stored in the cloud in organizations worldwide from 2015 to 2022"

- How long (and saturated) your backup window is

- How long your recovery window is

- How "ready" you truly are to recover from a cyberattack or other disruption event

The result for many orgs is a dangerous slowing down of backup and recovery, an increase in cloud storage costs, and a lack of true disaster recovery readiness.

There are two premises I want to ask you to keep in mind while you read this eBook:

The first is: *No one's concerned about how fast your backups are, but <u>everyone</u> cares how long recovery takes.*

Add the second is: *Your organization's budget dedicated to cloud-based storage isn't going to grow exponentially like your data is.*

What this means for you is that even as your organization's data grows, the expectation will always be that you can a) still back everything up, b) recover it quickly, and c) do it without significantly raising costs.

That's a pretty tall order – one you already likely realize you eventually will not be able to meet.

What's needed is to determine how to best leverage the cloud for backup, disaster recovery and data retention, keeping cloud storage costs under control, while improving the speed of recovery.

Optimizing Backups by Optimizing Backup Data

Your focus is on speed of execution for accurate backup and recovery, while your organization's focus is on doing so within or under budget. In the end, you both need to win. So, how can you go beyond the backup and recovery you currently have in place and meet both goals?

The answer is in *optimizing the backups themselves*.

There are three methods of optimizing backups you should consider:

- Data Deduplication – This reduces the size of the overall data being backed up and stored in the cloud by eliminating redundant data and storing only unique data sets.

- Data Compression – This also reduces the backup data size by encoding the data using fewer bits than the original data set being backed up.

- Storage Tiering – This is a strategy where you optimize the use of cloud storage classes to save costs.

Let's dive into each of these a bit more, paying close attention to not just how they can be used to optimize your backups, but also to where in the backup process you should be performing each type of backup optimization.

Data Deduplication

Deduplication uses an index to track unique blocks or files being deduplicated. For each block of data or file within a given backup set that is found to be unique, the index is generally updated with a hash (for blocks) or a pointer (for files). When a duplicate block or file is found, it is replaced by either the hash (blocks) or a file stub (files) to reduce the overall size of the backup data being stored. Deduplication can achieve (depending on how it's implemented) a space savings well into the 90%+ range.

There's also *variable-block* deduplication. Basic block-level deduplication assumes a consistent block size (e.g., 4K, 8K, 16K, 32K). Keep in mind that the type of backup data being deduplicated may be better suited for a larger or smaller block size in order to find more matches and achieve the greatest size reduction. So, variable-block deduplication identifies patterns in the data, adjusts the block size used, and can even align block start and end points to get even more space savings.

While the percentages of savings achieved vary between deduplication technologies, the basic idea around how much space is saved is:

File < Fixed-Block < Variable Block

Data deduplication can be processor-intensive so there are some considerations to think about in your approach:

- Source vs. Destination – Does the deduplication occur *before* it's backed up to the cloud (*source*) or once it's in the cloud (*destination*). Generally, because system performance (e.g., processor, memory, etc.) is improving at a much faster rate than network speed, available bandwidth, and latency, source-side deduplication is considered better for cloud-based backup and recovery.

- Job Level vs. Global – Is deduplication being done only for the data within a single backup job (e.g., a few VMs) or across all of

your backup jobs (e.g., your entire enterprise). The general consensus is that global deduplication will result is the greatest reduction in backup data sizes.

- Deduping SaaS Data – Assuming your backup strategy includes your Microsoft 365 and Salesforce instances (and other key cloud applications), you should be deduplicating that data as well.

- Software-Defined vs. Appliance-Based – While deduplication does exist in most backup solutions, many organizations want a solution dedicated to deduplication. Some utilize physical appliances, which then means you need to worry about hardware issues, refreshes, etc. Software-defined deduplication runs as a VM that can be hosted on-prem or in the cloud, where as much or as little CPU and RAM can be devoted to the VM as desired to achieve the deduplication performance needed.

There's often a question about the difference between *deduplication* and *incremental forever* backups. While both (in essence) are only backing up the changed data, incremental backups only benefit the current backup job, whereas global deduplication benefits every backup.

Compression

If you've been in IT long enough, you already get compression. In a backup context, the goal is to compress the data before it's sent to the cloud to reduce the amount of data traversing between where your workloads and your cloud backups exist. It's necessary to balance the desired resultant compression ratio with the latency it creates in the backup and restore processes, due to its' processor-intensive nature.

For reference, a compression ratio under 1:10 (that is, the *compressed size* vs. the *original size*) is considered to be good performance. Different data

types compress at different ratios, and anything above 1:10 (e.g., 1:12) is considered to be excellent.

The very nature of compression being processor-intensive should raise the question of whether you should compress at all. The short answer is yes, with the caveat that you want the benefit of compression as long as it isn't slowing down your backup and recovery.

Storage Tiering

Both of the previous optimization methods can be applied to on-prem and cloud-based sources and targets. But cloud storage has one more option that you may or may not be taking advantage of. Every major cloud provider has several tiers of cloud storage they offer. The tiers of storage are generally designed to represent how quickly you need the first byte of data during recovery – your critical backups are needed instantly, whereas recovering something from eight years ago probably isn't as pressing to see a restore start immediately. There are other factors that come into play with each tier (e.g.,

availability, storage duration minimums, minimum storage charges per object). As you shift toward the slower tiers, there is lower storage pricing per GB.

For example, at the time of this eBook being written, AWS has seven cloud-based storage tiers, Azure has four, and Google has four.

You might wonder why you don't just use the cheapest tier of storage always. Remember the first premise I mentioned to keep in mind throughout this eBook? In the middle of a recovery, you want it measured in minutes, right? So, if you use the "coldest" tier of storage on AWS, for example, the first bye latency (that is, how long you should expect before you get byte one of your backups to recover) is simply listed as "hours".

Instead, you should be optimizing backup costs in the cloud by having a policy-based, automated means of moving retained backups to colder tiers of storage to save on costs. For example, when a backup data set reaches three or six months old, it's automatically moved to the next lower tier of storage.

Remember, you can always recover from any tier; the issue is balancing how quickly you need it with how much it's costing you to retain it.

Practically Optimizing Your Backup Data

First, let's start with the business objectives you should be translating into technical requirements to get the most from your backup data. There are two basic requirements that every organization would have around backup data:

1. Be able to recover any part of the environment as quickly as possible

2. Do so while keeping the cost of storage as low as possible

In short, your goal is to balance recovery speed with cost.

So, what can you do practically to achieve this?

I'll start with the things you can likely do this moment, and then move onto the things that may cost money, take more time, etc.

Use Tiered Storage

I'm going to assume you have some data retention definitions that establish how long the backups of certain workloads need to be kept and made available long-term – including how quickly they can be retrieved (remember, different storage tiers have different first byte retrieval times). Do the research to understand how many tiers your cloud storage provider offers, what the cost differences per GB are, what's the first byte latency, and then map all this to the workloads and operational data your backup data sets represent.

In the end, you should have a better-defined data retention strategy that includes not just duration, but storage tiers to be used *over time*.

To take advantage of tiered storage, data has to be moved between tiers. Worst-case scenario is you use your cloud's native tools (whether a command line, control center, etc.) but do it – those backups of your AD that you're keeping for well-past six months "just to be safe" don't need to be sitting in more costly cloud storage. Some cloud providers offer storage management; that's a solid option. Many enterprise backup solutions also have a tiered

storage management capability built in. If this applies to you, take advantage of it there.

With the right solutions in place, it's also possible to "tier" on-premises backups and cloud backups. Your most recent backups remain on-prem for fastest recovery (assuming we're talking about on-prem workloads), with older backups being "tiered" (read: moved) to cloud-based storage.

Deduplicate the Data

Some backup solutions either include deduplication or have it as an add on. So, you can start there, but keep in mind that there may be limitations in how much benefit you get. For example, if your backup is deduping at a VM level, the best-case scenario is if you build, say, all 100 of your servers from the same base image, there would be a massive opportunity to dedupe all the commonality across them. But that's not likely the case, so read the documentation to better understand at what level any built-in

deduplication capabilities function – and how they are really going to benefit you.

Ideally, the variable-block deduplication is going to give you some solid savings, so be looking for that as an option. But whatever deduplication functionality is available to you, as long as it's implemented in a way that balances overall backup and recovery performance, I'd recommend taking advantage of it.

Compress the Data

I'm going to be a stickler here. Compression can be a CPU-intensive function and despite most backup solutions providing some degree of compression to their backups, it's not their specialty. Keep in mind that a backup solution's focus first and foremost is to back up and recover the data – which means, when it comes to compression, it will help, but it's not necessarily the best.

Be careful with compression! In some backup solutions, achieving just 10% more compression of the backup data can increase the CPU usage *by ten times*!

Definitely use the compression, but realize you're not going to get the full benefit of what can be accomplished today with backup solution-based compression.

Do All Three

This is where you can really benefit in both reducing storage costs and increasing backup and recovery speeds. But, there's a rule of marketing I have held onto for a number of years – "if your product does too many things, it doesn't do all of them well." Practically speaking, your backup solution can't be *best of breed* at backup and recovery and at compression, deduplication, and tiered storage management, right? *Right.*

You might consider looking for a solution that seamlessly sits on top of your backup solution to handle the compression, deduplication, and storage management so that your backup solution can be optimized to ensure backup and recovery efforts work at the highest levels of efficiency and accuracy. It's going to mean spending some budget on a third-

party solution, but the assumed improvements in data reduction and optimized storage will likely more than pay for the cost of letting a best of breed solution handle it instead of adding more overhead to your backup solution.

The Big Takeaways

Your backup and recovery efforts are inevitably going to become more taxing as data will undoubtably grow in the future. Because the organization will never have an unlimited budget for cloud storage, it's necessary for IT to leverage ways to optimize the data to increase the speed of backup and recovery while reducing the cost of retaining *and* maintaining the resultant backup data in the cloud.

By optimizing backups using deduplication, compression, and tiered storage, organizations proactively ensure they are using the least expensive cloud storage using the smallest amount of backup data possible, while simultaneously making certain that backup and recovery efforts occur as quickly as possible.

Quest® QoreStor®

Make your backup solution better!

Quest QoreStor will make whatever backup solution you use even better. Exponential data growth continues to be a major challenge but now you can fully optimize your backups and recover faster in both the data center and in the cloud.

Accelerate backup and recovery

Reduce backup storage requirements and cost

Connect to the cloud for backup, disaster recovery and long-term data retention

Gain ransomware protection via immutable backups.

Learn More

Quest®

Where Next Meets Now.

The rate at which corporate data is growing is putting a strain on IT's ability to cost-effectively back it up and recover it quickly when using cloud storage. I'll cover what's the root cause of the problem and provide four ways to optimize backups to lower storage costs and speed up backup & recovery.

About Nick Cavalancia

Nick Cavalancia is a 4-time Microsoft Cloud and Datacenter MVP, has over 25 years of enterprise IT experience, is an accomplished consultant, speaker, trainer, writer, and columnist.