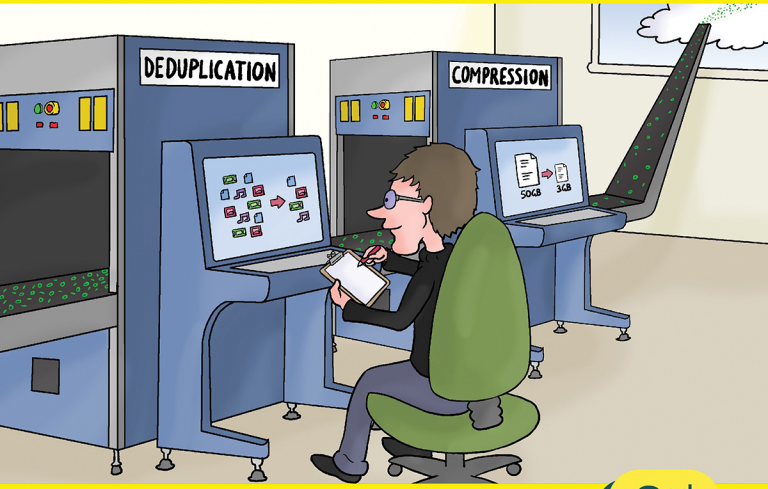
ConversationalGeek®

Conversational
Backup Optimization for the Cloud

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



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

2nd MINI Edition

Sponsored by

Quest®

Sponsored by Quest®

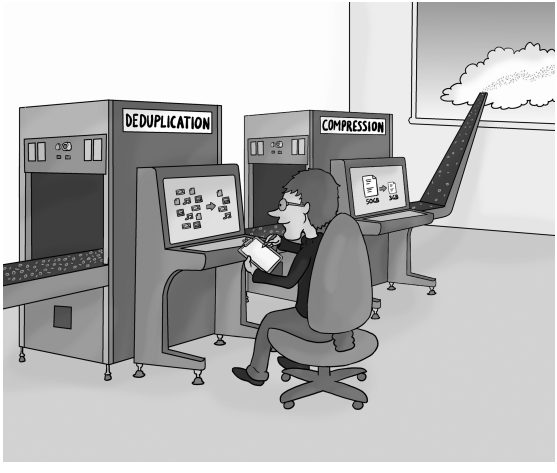
Quest Software delivers cybersecurity, data performance, directory and identity, and modernization solutions that simplify enterprise IT management. Quest solutions protect and empower users and data, streamline IT operations and harden cybersecurity from the inside out. Quest has the experience to help you manage and monitor your databases, take control of your hybrid environments and secure them all with security solutions that ensure you are ready for the next threat.

We're Quest – Where Next Is Now.

For more information, visit www.quest.com

Conversational Backup Optimization for the Cloud (Mini Edition)

by Nick Cavalancia



ConversationalGeek®

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):	Blathnaid O'Brien
	Adrian Moir

The “Conversational” Method

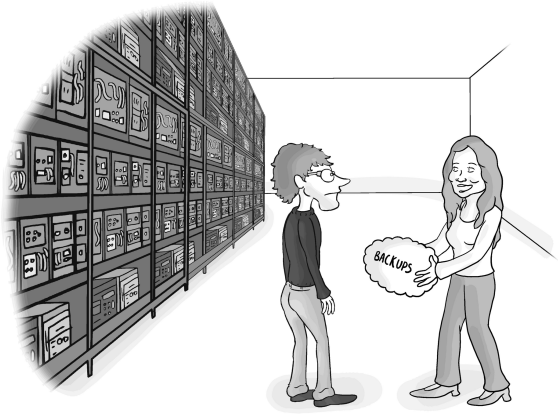
We have two objectives when we create a “Conversational” book. First, to make sure it’s written in a conversational tone so that it’s fun and easy to read. Second, to make sure you, the reader, can immediately take what you read and include it into your own conversations (personal or business-focused)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

Data, the invaluable asset of business in the 21st century, is surging at an exponential rate due to digitization. This rapid increase in data volume often leaves businesses struggling to keep pace with the need to securely and effectively back up all of their data. The total global data storage is projected to surpass 200 zettabytes by 2025[1], encompassing data on public and private clouds, utilities, personal devices like PCs and phones, and IOT (Internet of Things) devices.

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

In 2022, 60% of corporate data was stored in the cloud[2] - and that number has continually grown, with 93% of organizations believe their use of cloud storage will increase in 2024[3].

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

---

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

[3] Wasabi, *Global Cloud Storage Index* (2024)

- How much you pay for storage
- 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 Backup Data = Optimized Backups

How do you optimize backups so that you can meet the goals mentioned? There are three main methods you should explore:

- Deduplication –This technology reduces the size of your backup by eliminating

redundant data and storing only unique data sets.

- Compression- This reduces backup size by encoding the data using fewer bits than the source data being backed up.
- Storage Tiering – A strategy of optimizing the use of different cloud storage classes to optimize spending.

Let's examine each of these methods a bit deeper to see how to use them and where they yield the most benefit in the backup process.

Data Deduplication

Imagine your data was a physical library of books. Some of the books are duplicates, like the spreadsheet that everyone has a copy of. These duplicates take up unnecessary space. Data deduplication is a way to clear the clutter. There are two main methods of deduplication: basic and block.

- Basic Block Deduplication – In this method, data is broken into smaller equal-sized blocks (usually 4KB to 128KB) that are then

compared to each other. The process begins by scanning these blocks to identify duplicates. If two or more blocks are identical, only one instance is retained, and subsequent occurrences are replaced with pointers or references to the original block. Basic block data works well with structured data where the data has the same 'size.' Depending on how it's implemented and the data being deduplicated, deduplication can save space into the 90%+ range.

- Variable Block Deduplication—This method takes a more adaptive approach. Algorithms determine the optimal size for each block. If the chunk of data is unique, it gets its own block. If it's similar to existing blocks, it shares a block. This can lead to more efficient storage utilization as some data needs big blocks, and some will fit into smaller blocks. Variable Block Storage really

shines with unstructured data because it can adapt to the data's natural patterns.

The percentage of savings will vary between deduplication technologies, but a good rule of thumb on the space-saving is:

File < Fixed-Block < Variable-Block

Because deduplication can be processor-intense, how you approach the solution can significantly impact your time to back up and recover. Here are a few decision points to consider while planning:

- *Source or Destination? – This one is pretty straightforward. Do you deduplicate the data before you send it to the cloud (source) or after it's uploaded (destination)? There are two very good reasons to do source-side deduplication. The biggest is that system performance (CPU, Memory, etc.) improves faster than networking speed, available bandwidth, and latency. Nearly as important is the "ingest" or upload cost of your cloud*

storage. Simply put, it costs less to upload one terabyte than to upload ten terabytes of data.

- *Job vs. global level—Is your deduplication "comparison set" restricted to a single job, or is it global across all of your backup jobs? The larger the sample, the more effective the deduplication, so it's generally considered best practice to go with the global approach.*

- *Software Defined vs Appliance-Based – It's true that most backup software offers some deduplication, but most organizations prefer a dedicated deduplication solution. These usually are either hardware-based or virtual appliances. Hardware appliances bring their own complications like hardware lifespan, warranties, power consumption, rack space, etc. Virtual appliances run like a VM in your virtual environment, either on-prem, in the cloud, or both. This approach allows you to*

tune the CPU and RAM allocation to achieve the data deduplication performance needed for your environment.

Don't confuse incremental backups with deduplication. Incremental backups only reduce the size of the existing job; deduplication reduces the size of every job.

Compression

This one has been around for a long time as a way to squeeze more data onto a disk. In a backup to the cloud scenario, it's a good idea to compress the data before it's sent to the cloud for many of the same reasons to deduplicate before sending the backup data to cloud storage. Specifically, the less data that has to traverse the network, the faster your restores will be. However, a balancing act needs to happen as compression is also processor-intensive. Attempting too much compression will introduce latency into your backup and recovery process.

For reference, compression ratios under 1:10 (compressed size vs original size) are considered good performance. Different types of data will compress differently, largely depending on whether the file type has compression built in, like JPG files. Anything better than 1:10 (like 1:12) is considered excellent.

Deduplication and compression can be used together!

It's like packing for a trip. You don't take three of the same shirts (dedupe) and then put the shirts you do take into a vacuum packing cube(compression).

Tier the Storage!

Both deduplication and compression methods can be applied to on-prem or cloud-based sources and targets. However, cloud storage has other features that you may or may not be fully utilizing.

Every major cloud provider offers several tiers of storage. These tiers are usually designed around how quickly you need the first byte of data during a recovery operation. Recent critical backups are

needed instantly (or as close as possible), while recovering something from several years ago can likely wait a bit for the restoration process to begin. Other factors that come into play between the different tiers are availability, storage charges per object, and storage duration minimums. Generally, as you move toward the slower, less easily accessible tiers, the price per GB drops significantly.

So why not always use the cheapest tier? Think back to RTO (Recovery Time Objective). Remember how we wanted that number to be measured in minutes? The coldest tiers of cloud storage have the time to recover the first byte of data listed as "hours"—one or two, eight or even more.

A better solution is to work the tiers with a policy-based automated method of moving older backups to slower, colder tiers to reduce costs. This can be extended to a hybrid solution where backups are kept locally for 48 hours, then moved to the fast (expensive) Cloud tier, where it stays for thirty days. Then, it's moved to the middle tier for three to six months, after which it moves to the lowest, cheapest tier for long-term storage. Don't forget that you can always pull from any tier; the

difference is how fast you can access the data. The balancing act of cost vs. speed is unique to every organization and even different data sets within an organization.

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.

Update Your Data Retention Policy

I'm assuming that you have written retention policies dictating how long backups of certain workloads or data types should be kept and available. Do your policies specify a time frame for retrieval? They should. This ensures that expectations are aligned between IT and the rest of the business. Nothing ruins an evening like a 5 PM request from Legal to have all of the documents related to a job or order from five years ago available first thing in the morning.

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 and anomaly detection

Learn More

Quest®

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 30 years of enterprise IT experience, is an accomplished consultant, speaker, trainer, writer, and columnist.