# Backup vs. Archive Why It's Important to Know the Difference

**Backing up and archiving data have distinct functions, and not recognizing that it's important to have both can lead to access problems and even legal troubles.**

It's just shy of saying that data on a RAID array doesn't need to be backed up. The good news is that the differences between backup and archive are quite

What is backup?

Backup is a copy of data created to restore said data in case of damage or loss. The original data is not deleted after a backup is made.

Examples of backups include a nightly backup of all files on your laptop or desktop, or all your photos on your iPhone being copied to iCloud in case you drop your phone. We also backup file servers (unstructured data) and databases (structured data). A backup might focus on the data, as in a database dump, the operating system of the server as in a bare-metal backup, or on both as with backup of VMware .VMDK files.

The definition of backup really comes down to purpose, and the purpose of a backup is always the same: to restore data if something happens to it. For example, a RAID

6 array might have a triple-disk failure, and all its data will need to be restored. Someone might accidentally or maliciously delete one or more VMs in your VMware, Hyper-V or AWS EC2 configuration, and they would need to be restored. You might one day realize every file in your organization has been encrypted by a ransomware package. Without a good backup system, your choice would be to pay the ransom. With a good backup system, you could figure out the source of the ransomware, stop it, then restore all your data – without ever paying the hacker.

What is an archive?

An archive is a copy of data created for reference purposes. Although not required, the original is often deleted after an archive is made.

Where the purpose of a backup is to put something back to how it looked (usually) yesterday, an archive can serve multiple purposes. The most common purpose is to help you find some data from a long time ago. It could be single file that had a really important item in it, such as a contract a customer signed several years ago. It might be a related group of data, such as all the structural drawings of the building that just collapsed. Or it might be all the CAD drawings of the widget your company used to make that went out of style but is now back in style.

Another related data set might be all emails and/or files that can prove a given point. Perhaps an employee believes they were given permission to moonlight, and then was fired for doing so. Their lawsuit might issue an electronic discovery request asking for all emails to and from them that contain the words moonlight, after-hours or the name of the company they were going to moonlight for. Someone else might be trying to prove a hostile work environment and want to see all emails from a particular set of managers that contain a certain list of words that we do not need to list here.

An archive is what would help you accomplish all of these tasks. You might have an archive of every sales order, quote or contract your company has ever given. You might keep current contracts and orders online, but you keep all of them in the archive, which should have an index to let you retrieve orders and contracts via the content of those orders. You also might have an archive of every email ever sent or received by your company.

Some email archive systems can purge from the email server emails that have been archived, are bigger than a certain size, and/or haven't been accessed in over n days. This helps keep the email system lean, saving on computing and storage resources, and making it easier to backup. That might even be the purpose of that archive, if you're not required by law to keep all emails.

Restore vs. retrieval

Even if the purpose of an archive is to save space on primary storage, it needs to be able to perform a retrieval vs a restore if it is to be called an archive. Backup systems restore and archive systems retrieve.

When you restore something, it is typically a single file, server or database. When you retrieve something, it's usually a collection of related data, that may or may not have been stored on the same server or even in the same format. A restore is also done to a single point in time, such as restoring a database to the way it looked yesterday. A retrieval uses a range of time, such as all emails for the last three years.

Restores require you to know a lot of about where the file or data was when it was backed up; otherwise, you can't find it. You need to know the name of the server it was on, the database or directory it was in, the name(s) of the file or table you want back and the date when it was last seen. Retrievals have none of that information; they just know they need all the files or records that match a set of parameters. Give me all files or emails that were created in the last three years that contain a particular phrase or were authored by a particular person.

Why the difference matters

Many people try to use their backup system as an archive system, meaning they keep their backups for many years – or even forever. The first time you get a real retrieval request, you'll find how difficult it is to perform a retrieve from something that is mean to do restores. This will make the retrieval take much, much longer – potentially months instead of minutes – and cost much, much, more – millions instead of a few dollars.

If the retrieval is for an electronic discovery request from a lawsuit, and you are unable to satisfy it in a timely manner, you run the risk of the judge issuing an adverse inference instruction. You've taken six months to satisfy what they know to be a simple request, and you're nowhere near complete. The judge infers you're trying to hide something, and they say that to the jury. You just lost the case. The most infamous example of this was the Morgan Stanley lawsuit where they lost billions in this exact scenario.

Don't use your backups as archives. If you have a long-term storage need, investigate an actual archive system. There will be an upfront cost, but it will be worth it in the long run.