Protecting Unstructured Data at Scale with Veeam Data Platform

Contents

Introduction

With the Veeam Capacity Pack licensing, you are equipped with powerful tools to protect your unstructured data. This solution brief will guide you through licensing unstructured data, provide configuration instructions for protecting it, and share best practices to help you maximize your Veeam solution. This document aggregates information from various sources to create a straightforward and easy-to-follow guide. We hope this document provides you with the essential information you need to begin protecting your critical unstructured data.

> **NOTE:** Should you have further questions or need technical assistance, you can reach out to your Veeam representative or open a support ticket.

How Unstructured Data is Licensed with Capacity Pack Licensing

Unstructured data is licensed per terabyte (TB) of non-deduplicated and uncompressed front-end source data. This licensing model provides flexibility in managing and protecting large volumes of unstructured data.

There are different Capacity Pack editions offered alongside their respective Veeam Data Platform editions:

- **Premium, Advanced, or Foundation edition:** These editions are suitable for larger NAS and Object Storage unstructured data backups and are generally greater than 250TB. Licensing is offered in increments of 1TB for these editions.
- **Essentials:** Designed for smaller unstructured data backups, this edition is available in increments of 1TB with a minimum of 5TB and a maximum of 50TB.

When a Capacity Pack license is present, all unstructured data will be protected by the Capacity Pack license, and Veeam Universal Licenses (VUL) will no longer be used to protect any unstructured data.

By diving into the details and examples provided here, you'll be able to manage your license capacity effectively while ensuring comprehensive protection for your unstructured data.

Calculation of License Capacity

Veeam Backup & Replication rounds the protected amount of data for each unstructured data source down to the nearest TB. It also keeps track of the capacity consumed by protected data sources by recalculating the license consumption every 30 days or during each unstructured data backup job run. Data sources processed within the last 31 days are considered protected, and each protected data source consumes capacity from the license scope.

Here's how the licensing would be calculated for different volumes of unstructured data:

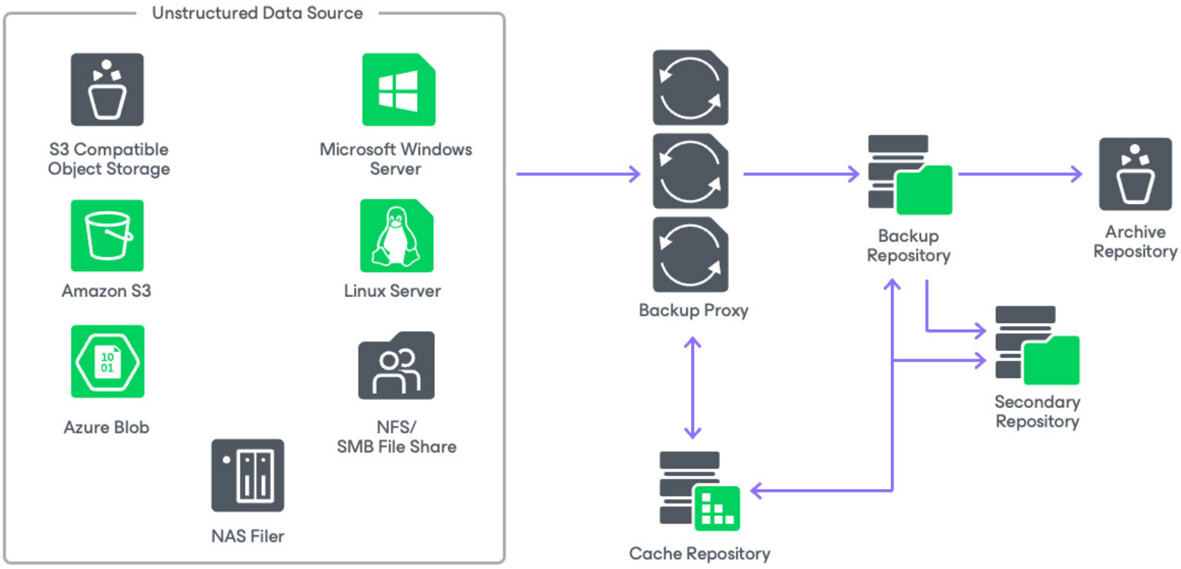
- **Example 1:** If you have a file server with 2.7 TB of non-deduplicated and uncompressed data, Veeam Backup & Replication will round it down to 2 TB, consuming two license units.
- **Example 2:** If you have a file server with 1.5TB of deduplicated and compressed data, Veeam Backup & Replication will protect the non-deduplicated and uncompressed data size. If the non-deduplicated and uncompressed data size is 2.9 TB, then it will be rounded down to 2TB, consuming two license units.
- **Example 3:** For two separate data sources, one with 1.3 TB and another with 0.9 TB, Veeam Backup & Replication will round each down separately to 1 TB and 0 TB, respectively, consuming a total of one license unit.
- **Example 4:** If a NAS device contains multiple file shares, such as \\root\share1 with 2.4 TB and \\root\share2 with 3.8 TB, Veeam Backup & Replication will add the protected amounts, rounding the total of 6.2 TB down to 6 TB, consuming six license units.

> **NOTE:** Unstructured data sources that are backed up to tape or object storage are excluded from capacity consumption calculations. Additionally, specific file extensions related to Veeam backup files will not consume license capacity.

Overview of Unstructured Data Backup

Backup Infrastructure for Unstructured Data Backup

Unstructured data backup requires a few unique components to function properly. Let's go over these different components and what they are used for below.



Unstructured Data Source

Unstructured data, by definition, is data not organized into predefined data models or databases. While this data can follow user-defined structures, the data is generally comprised of large amounts of individual files. Some examples of the types of files that could make up unstructured data would be text-based documents, images, videos, audio files, spreadsheets, and much more.

Many different unstructured data sources are supported for backup and restore with Veeam Backup & Replication.

For more detailed information, please visit: Unstructured Data Source.

General Purpose Backup Proxies

A general purpose backup proxy is responsible for transferring data between an unstructured data source and a backup repository. It processes jobs and handles the traffic associated with backup and restore operations.

> **NOTE:** By default, this role is assigned to the backup server itself, which is adequate for smaller installations where all components are situated within the same network segment. For larger installations with more extensive workloads, it is advised that you assign the role of a backup proxy to a dedicated server.

To enhance the performance of multiple concurrent tasks, you can use multiple backup proxies. In these scenarios, Veeam Backup & Replication will distribute the backup or restore workload among available backup proxies on a per-task basis while considering proxy connectivity and their current load.

To reduce network load during backup operations, position the backup proxy closer to the source file share or storage within the computer network. Ideally, they should be only one step away from each other.

For detailed system requirements, please visit: General-Purpose Backup Proxy System Requirements.

Cache Repository

A cache repository in Veeam Backup & Replication stores temporary metadata to reduce load during backups. It tracks changes between sessions to enable quick incremental backups from unstructured data sources. When using an object storage repository as the target, the cache repository also maintains active metadata.

You can assign a backup repository as a cache repository within Veeam's infrastructure when adding an unstructured data source. For optimal performance, position the cache repository close to the backup proxy, ideally on the same machine.

> **NOTE:** Deduplicating storage appliances cannot serve as cache repositories.

For detailed system requirements, please visit: Cache Repository System Requirements.

Storage Repositories

A backup repository is the primary storage location where Veeam Backup & Replication stores backed up unstructured data. Backups in the backup repository can be used to restore an entire file share to a specific state or restore individual files as needed.

If specific files need to be retained for a longer period, potentially cheaper and more dense devices can be used for archiving purposes. Veeam Backup & Replication will move backup files and metadata files from your primary backup repository to an archive repository following specified retention.

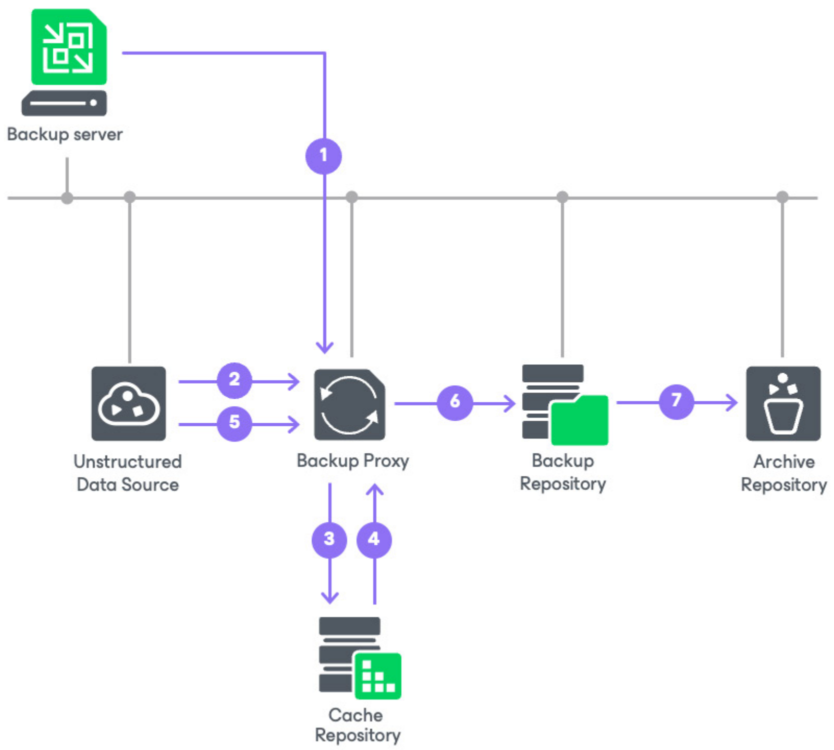
Secondary repositories can be configured to store a copy of the unstructured data backup in a different location for redundancy purposes. The secondary repository can have its own retention policy and encryption settings.

> **NOTE:**
> - By default, use of an archive repository is disabled, and after the retention period for your backup repository is over, backup files are deleted.
> - By default, no secondary repository is configured.

For a detailed list of storage types that you can leverage, please visit: [Storage Repositories Table](#).

How Unstructured Data Backup Works

Let's look at how Veeam Backup & Replication handles the backup process for unstructured data. In a high-level view, this operation happens in seven steps. Each of these steps are listed with a visual diagram below:



1. When a backup job starts, it will assign a backup proxy.
2. The proxy scans source files and folders and creates a CRC tree.
3. The CRC tree is sent to the cache repository.
4. The cache saves the CRC tree and compares it with the previous one.
5. If there are changes, the cache instructs the proxy to read the new data.
6. The proxy reads new data and creates data packages. These packages, which contain 64MB of backup data files and metadata, are sent to the target repository.
7. Veeam checks file versions against retention settings and moves data to the archive if needed.

Unstructured Data Retention

Unstructured data retention is handled a bit differently than a standard backup job since it is not keeping one singular backup file with data inside it. Instead, it is keeping files as objects in separate VBLOB files (up to 64MB each) alongside the metadata which details the original objects on unstructured data sources and their structure. This means that retention is handled on each file and its subsequent file version changes.

Retention for unstructured data backup is defined as the number of days or months to retain file versions, and the number of active or deleted file versions to keep. The default is all file versions are kept.

The basic rules for retention are as follows:

- Retention policy settings apply against file version changes. Unchanged files with no file version changes will remain in the primary backup repository indefinitely, regardless of retention policy if no archive repository is configured.
- When a file is deleted from the source, the backup repository marks its version as deleted and continues retaining and deleting according to your set retention period.
- Excluded files are not included in the backup process and therefore are not included in retention.
- Retention is applied to file versions by first following the file version limit set before applying time-based retention.

The archive repository has its own defined retention policy for an unstructured data backup job. Archive retention also allows data to be copied alongside the primary backup repository to make redundant copies, file extension inclusion or exclusion, and limits on the number of file versions to keep.

The archive retention rules are as follows:

- File versions that are older than the primary backup repository retention but lie within the defined archive retention period are moved to the archive repository. Newer file versions that are still within primary backup repository retention are retained in the primary backup repository.
- File types that are excluded from archive retention would be deleted from the primary backup repository once they're older than the defined retention period, instead of moving to the archive repository. Unexcluded file types would continue to move to the archive repository where applicable.
- Retention is applied to file versions by first following the file version limit set before applying time-based retention.

When using an immutable repository like a hardened repository, an immutable deduplicating storage appliance, or an immutable object storage repository, data cannot be altered or removed until the immutability period ends.

The immutability retention rules are as follows:

- The immutability period set always trumps any other retention period due to the lock on files that's present until the immutability period expires.
- Immutability is marked on each individual file version that's backed up.
- File versions are immutable for the length of retention plus the immutability period.
- When a file is deleted from the source, the file is marked as deleted and is then retained for the length of the defined retention period plus the immutability period before being fully deleted.
- For immutable object storage repositories, the Block Generation period (up to 10 days) is automatically added to the immutability period to extend the overall immutability timeframe.

> **NOTE:** When backing up unstructured data to an immutable repository, Veeam recommends that you use no more than 14 days of immutability due to additional storage usage for each day of added immutability.

For examples of how retention is applied, please visit: Unstructured Data Backup Retention Scenarios.

Sizing for Unstructured Data Backup

When it comes to protecting unstructured data with Veeam Backup & Replication, it's essential that you size your components properly to ensure efficient and reliable backups. Sizing the components and storage requirements for protecting your unstructured data is easy using Veeam calculators. Veeam calculators give you the option to calculate the resources needed either overall or per workload.

To properly calculate sizing requirements, please visit: Veeam Calculators.

For more information about minimum system requirements, please visit: System Requirements.

Gather Share Information

Proper component sizing requires some information to be gathered. To keep sizing calculations simpler, we will focus on gathering overall estimates for your unstructured data to be protected. Once this information is gathered, you can feed it into Veeam calculators, listed above, to get required resource information.

General Overview Information

Total source data size (non-deduplicated and uncompressed):	
Number of source shares:	
Average read speed per share (in MB/s):	
Average max allowable read speed per device (in MB/s): *This is the read speed that's allowable without affecting day-to-day operations	
Number of files (in millions):	
Daily change rate percentage:	
Types of files (e.g., multimedia, images, documents): *This is used to estimate the compression of the backed-up data on the target repository	

Component Sizing Standards

This section is meant to give a quick overview of component sizing standards. These standards are from the Veeam best practices guide.

For more information about sizing best practices, please visit: Best Practice Guide.

> **NOTE:** The resource standards within this document may be different than other documents and guides since product performance is constantly being optimized in each release.

General-Purpose Backup Proxy

Sizing Standards	
1 source share:	2 proxy tasks
1 proxy task:	4 GB of RAM
1 proxy CPU core:	100 MB/s or 0.34 TB/hr read

Default Recommended Virtual Proxy	
CPU cores:	12 vCPUs
Memory:	16 GB of RAM
Proxy tasks:	4 tasks

> **NOTE:** Adding more CPU cores to the general purpose proxy increases the rate at which it can read data. While this can be beneficial, it could lead to oversaturation of the source device, thus causing production slowdowns.

Due to this, it's best practice to work toward fine-tuning your general purpose proxy deployment.

Cache Repository

Sizing Standards	
1 source share:	1 cache repo agent
1 cache repo agent:	4 CPU cores
1 cache repo agent:	8GB of RAM
Storage needed for disk based repositories:	Max 4GB
Storage needed for object based repositories:	5% of the source data size

> **NOTE:** A source share is a share that's added into the inventory section, not subfolders.

Target Repository

Sizing Standards	
CPU cores:	½ general purpose proxy CPU cores
1 CPU core:	4GB of RAM
Disk storage space:	Calculated based on full backups, incrementals, metadata, and workspace requirements.

> **NOTE:** The storage requirements for the target repository will be different for different types of repositories due to how metadata operations are handled.

Configuration Steps

Configuring unstructured data backups with Veeam Backup & Replication is designed to be a straightforward and user friendly process. Regardless of your unstructured data source, setting up the components and configuring jobs to protect your data can be achieved in just a few steps. In this section, we will go through the basic steps to start protecting your unstructured data as more of a quick start guide. There are more options available that are not covered by this guide; this guide is meant to provide an overview.

Adding the Components

1. General Purpose Backup Proxy

> **NOTE:** You don't need to add an additional general purpose backup proxy if you use the pre-configured default option, which is the Veeam Backup & Replication server itself.

1. Navigate to "backup infrastructure".
2. Right-click on "backup proxies" and select "add proxy".
3. Select "general purpose backup proxy".
4. Select a Windows or Linux system from the list.

> **NOTE:** If you have not already added the Windows or Linux machine that you plan to use, then you will need to click the "add new..." button to add it.

5. For basic configuration, proceed through and finish the wizard.

For more information about all the options you have when adding a general purpose backup proxy, please visit: Adding General-Purpose Backup Proxies.

2. Cache Repository

NOTE:

1. If you have not already added the repository that you plan to use as your cache repository, then you will need to follow the steps below to add one.
2. The following can be used as a cache repository:

Direct attached storage (DAS)

- Windows Server
- Linux Server

Network attached storage (NAS)

- SMB (CIFS) Share
- NFS Share

3. Deduplication storage appliances cannot be used as cache repositories.

1. Navigate to "backup infrastructure".
2. Right-click on "repositories" and select "add repository".
3. Select the storage type you'd like to add.
4. For basic configuration, proceed through and finish the wizard.

For more information about all the options you have when adding a new repository, please visit: [Adding Backup Repositories.]

3. Unstructured Data Source

1. Navigate to "inventory".
2. Right-click on "unstructured data" and select "add unstructured data source".
3. Select the type of unstructured data you'd like to add.

NOTE:

1. The following can be used as an unstructured data source:

File shares:

- Microsoft Windows Server
- Linux server
- Enterprise NAS systems:
    a. NetApp ONTAP (including Amazon FSx for NetApp ONTAP)
    b. Lenovo ThinkSystem DM/DG Series
    c. Dell PowerScale
    d. Nutanix Files Storage
- NFS file shares
- SMB file shares

Object storage:

- S3-compatible object storage
- Amazon S3
- Microsoft Azure Blob storage, including Azure Data Lake Gen2

2. No support for backup from/restore to Azure Data Box Storage, Azure Stack Edge, AWS Snowball Edge Storage, or Veeam Data Cloud Vault.
3. If you have not already added the source that you plan to use, then you will need to click the "add new..." button to add one.

4. In the "processing" section of the wizard, define the repository you added previously as the "cache repository".
5. Note that "backup I/O control" defines how quickly Veeam can process your unstructured data source.
   - "Lower impact" reduces the number of proxies available to process each share, as well as the available threads for each proxy. When set to the far left, it will only use one proxy with a single thread.
   - "Faster backup" increases the number of proxies available to process each share, as well as the threads available for each proxy. When set to the far right, it will use any number of proxies defined for processing with up to 16 threads per proxy.
6. Finish the wizard and save the settings.

For more information about all the options you have when adding an unstructured data source, please visit: Adding Unstructured Data Source.

Configure Unstructured Data Backup Jobs

1. Navigate to **"home"**.
2. Right click on **"jobs"** and either select **"backup > file share..."** or **"backup > object storage..."**.
3. Specify the name of the job and click **"next"**.
4. In the **"objects"** section, select **"add..."** and then specify the unstructured data from your share you want to protect and click next.
5. In the **"backup repository"** section, select the backup repository where you'd like your unstructured backup data to be stored.

NOTE:

- Veeam NAS, file share, and object storage backups can go to any Veeam backup repository except a Veeam Cloud Connect repository.
- If the target of the backup is a Scale-out Backup Repository (SOBR), Capacity Tier and Archive Tier settings do not apply to unstructured data backups.

6. Define how many day's worth of file should be kept.
7. Optionally, select "**configure a secondary destination for this job**" if you'd like to create a backup copy of this data.
   - In the "**secondary target**" section, select one or more existing backup repositories where you'd like another copy of your unstructured backup data to be stored.
   - Select a secondary repository and click "**Edit...**" to a custom retention policy if necessary.
   - For steps on configuring a secondary target, please visit: Adding Backup Repositories.
8. Optionally, in the "**archive repository**" section, select "**archive file versions to the following archive repository**" if you'd like to archive this data.
   - Select an existing backup repository where you'd like your unstructured backup data to be archived.
   - Optionally, select "**archive recent file versions**" if you'd like to create a redundant copy of all backup data that's stored in the primary backup repository.
   - Optionally, select "**archive previous file versions for**" and define the amount of time you want data to be archived.
   - For steps on configuring your archive repository, please visit: Adding Backup Repositories.

NOTE:
- Archive repositories cannot be a SOBR or a repository with rotated drives.
- Amazon S3 Glacier and Azure Blob Storage Archive Tier are not supported for unstructured data backup.

9. Define your schedule.
10. Finish the wizard and begin protecting your data.

NOTE: Once backups are completed, it is highly recommended that you regularly test your backups to ensure the effectiveness of your data protection strategy.

For more information about all the options you have when adding an unstructured data backup job, please visit: Creating Unstructured Data Backup Jobs.

Conclusion

You should now be confident in your understanding of how Veeam Backup & Replication safeguards unstructured data and how it's licensing works. Most importantly, you should be able to successfully create and initiate a backup job. However, this is just the beginning.

Designed to meet the ever-evolving challenges of modern IT environments, Veeam Data Platform delivers secure, reliable, and rapid data backup and recovery to keep businesses running smoothly. Whether it's protecting critical workloads, preventing data loss, or ensuring seamless disaster recovery (DR), Veeam Data Platform provides the resilience and security that organizations need to stay ahead. Built for speed, engineered for protection, and trusted by millions of users worldwide, Veeam Data Platform's advanced technology protects data across cloud, virtual, physical, SaaS, and Kubernetes environments. With a future-proof approach, businesses can scale confidently, knowing their data is always protected, always recoverable, and always available when it matters most.

There are numerous additional resources available and a plethora of features to still explore. To learn more about Veeam Data Platform, please visit www.veeam.com.

About Veeam Software

Veeam®, the #1 global market leader in data resilience, believes every business should be able to bounce forward after a disruption with the confidence and control of all their data whenever and wherever they need it. Veeam calls this radical resilience, and we're obsessed with creating innovative ways to help our customers achieve it. Veeam solutions are purpose-built for powering data resilience by providing data backup, data recovery, data portability, data security, and data intelligence. With Veeam, IT and security leaders rest easy knowing that their apps and data are protected and always available across their cloud, virtual, physical, SaaS, and Kubernetes environments. Headquartered in Seattle with offices in more than 30 countries, Veeam protects over 550,000 customers worldwide, including 67% of the Global 2000, that trust Veeam to keep their businesses running. Radical resilience starts with Veeam. Learn more at www.veeam.com or follow Veeam on LinkedIn @veeam-software and X @veeam.

Learn more: veeam.com