Recover from Storage Node failures

StorageGRID 11.7

NetApp
April 12, 2024

This PDF was generated from https://docs.netapp.com/us-en/storagegrid-117/maintain/recovering-from-storage-node-failures.html on April 12, 2024. Always check docs.netapp.com for the latest.

Table of Contents

Recover from Storage Node failures . . . 1
- Recover from Storage Node failures: Overview . . . 1
- Recover Storage Node down more than 15 days . . . 2
- Recover appliance Storage Node . . . 3
- Recover from storage volume failure where system drive is intact . . . 25
- Recover from system drive failure . . . 40
- Restore object data using Grid Manager . . . 58
- Monitor repair-data jobs . . . 62

Recover from Storage Node failures

Recover from Storage Node failures: Overview

The procedure for recovering a failed Storage Node depends on the type of failure and the type of Storage Node that has failed.

Use this table to select the recovery procedure for a failed Storage Node.

Issue	Action	Notes
• More than one Storage Node has failed. • A second Storage Node has failed less than 15 days after a Storage Node failure or recovery. This includes the case where a Storage Node fails while recovery of another Storage Node is still in progress.	Contact technical support.	Recovering more than one Storage Node (or more than one Storage Node within 15 days) might affect the integrity of the Cassandra database, which can cause data loss. Technical support can determine when it is safe to begin recovery of a second Storage Node. **Note**: If more than one Storage Node that contains the ADC service fails at a site, you lose any pending platform service requests for that site.
More than one Storage Node at a site has failed or an entire site has failed.	Contact technical support. It might be necessary to perform a site recovery procedure.	Technical support will assess your situation and develop a recovery plan. See How site recovery is performed by technical support.
A Storage Node has been offline for more than 15 days.	Recover Storage Node down more than 15 days	This procedure is required to ensure Cassandra database integrity.
An appliance Storage Node has failed.	Recover appliance Storage Node	The recovery procedure for appliance Storage Nodes is the same for all failures.
One or more storage volumes have failed, but the system drive is intact	Recover from storage volume failure where system drive is intact	This procedure is used for software-based Storage Nodes.
The system drive has failed.	Recover from system drive failure	The node replacement procedure depends on the deployment platform and on whether any storage volumes have also failed.

Some StorageGRID recovery procedures use Reaper to handle Cassandra repairs. Repairs occur automatically as soon as the related or required services have started. You might notice script output that mentions “reaper” or “Cassandra repair.” If you see an error message indicating the repair has failed, run the command indicated in the error message.

Recover Storage Node down more than 15 days

If a single Storage Node has been offline and not connected to other Storage Nodes for more than 15 days, you must rebuild Cassandra on the node.

Before you begin

- You have checked that a Storage Node decommissioning is not in progress, or you have paused the node decommission procedure. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Decommission**.)
- You have checked that an expansion is not in progress. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Expansion**.)

About this task

Storage Nodes have a Cassandra database that includes object metadata. If a Storage Node has not been able to communicate with other Storage Nodes for more than 15 days, StorageGRID assumes that node’s Cassandra database is stale. The Storage Node can’t rejoin the grid until Cassandra has been rebuilt using information from other Storage Nodes.

Use this procedure to rebuild Cassandra only if a single Storage Node is down. Contact technical support if additional Storage Nodes are offline or if Cassandra has been rebuilt on another Storage Node within the last 15 days; for example, Cassandra might have been rebuilt as part of the procedures to recover failed storage volumes or to recover a failed Storage Node.

If more than one Storage Node has failed (or is offline), contact technical support. Don’t perform the following recovery procedure. Data loss could occur.

If this is the second Storage Node failure in less than 15 days after a Storage Node failure or recovery, contact technical support. Don’t perform the following recovery procedure. Data loss could occur.

If more than one Storage Node at a site has failed, a site recovery procedure might be required. See How site recovery is performed by technical support.

Steps

1. If necessary, power on the Storage Node that needs to be recovered.
2. Log in to the grid node:
   a. Enter the following command: `ssh admin@grid_node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

      When you are logged in as root, the prompt changes from `$` to `#`.+

> If you are unable to log in to the grid node, the system disk might not be intact. Go to the procedure for recovering from system drive failure.

3. Perform the following checks on the Storage Node:
   a. Issue this command: `nodetool status`

      The output should be `Connection refused`

   b. In the Grid Manager, select **SUPPORT** > **Tools** > **Grid topology**.
   c. Select ***Site*** > **Storage Node** > **SSM** > **Services**. Verify that the Cassandra service displays `Not Running`.
   d. Select **Storage Node** > **SSM** > **Resources**. Verify that there is no error status in the Volumes section.
   e. Issue this command: `grep -i Cassandra /var/local/log/servermanager.log`

      You should see the following message in the output:

      ```
      Cassandra not started because it has been offline for more than 15
      day grace period - rebuild Cassandra
      ```

4. Issue this command, and monitor the script output: `check-cassandra-rebuild`
   - If the Cassandra service depending on volume 0 is running, you will be prompted to stop it. Enter: **y**

     > If the Cassandra service is already stopped, you aren't prompted. The Cassandra service is stopped only for volume 0.

   - Review the warnings in the script. If none of them apply, confirm that you want to rebuild Cassandra. Enter: **y**

     > Some StorageGRID recovery procedures use Reaper to handle Cassandra repairs. Repairs occur automatically as soon as the related or required services have started. You might notice script output that mentions "reaper" or "Cassandra repair." If you see an error message indicating the repair has failed, run the command indicated in the error message.

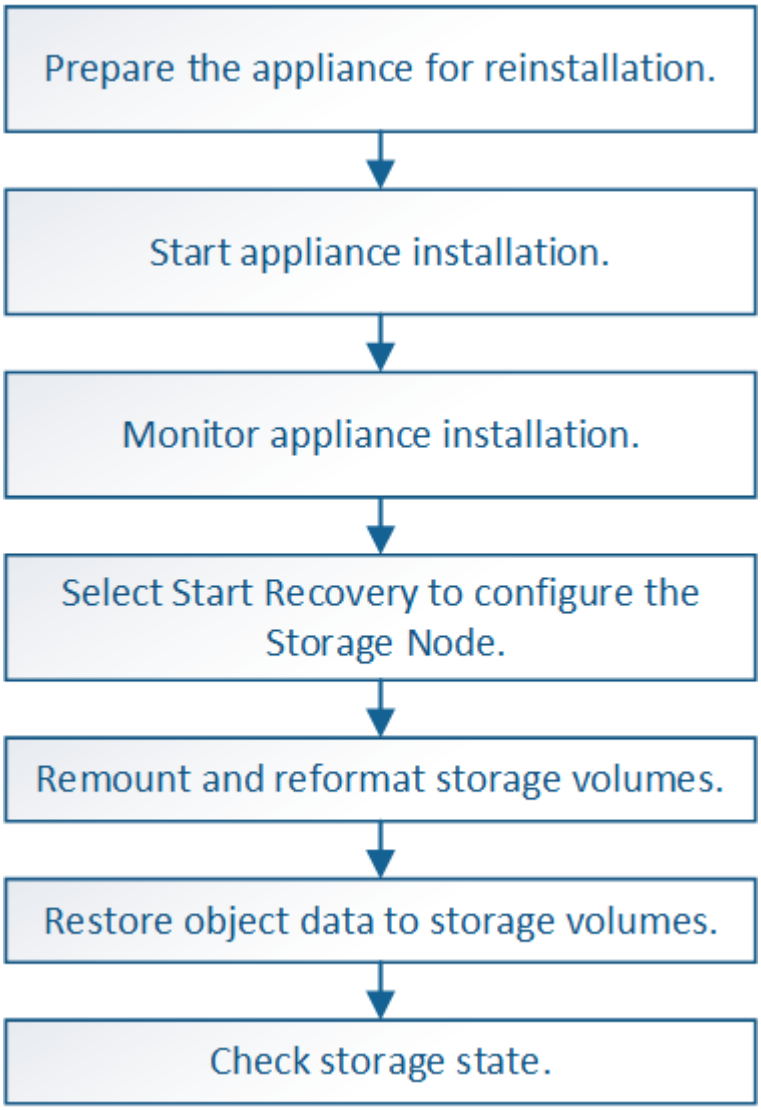
5. After the rebuild completes, perform the following checks:
   a. In the Grid Manager, select **SUPPORT** > **Tools** > **Grid topology**.
   b. Select ***Site*** > **recovered Storage Node** > **SSM** > **Services**.
   c. Confirm that all services are running.
   d. Select **DDS** > **Data Store**.
   e. Confirm that the **Data Store Status** is "Up" and the **Data Store State** is "Normal."

Recover appliance Storage Node

Recover appliance Storage Node: Workflow

The procedure for recovering a failed StorageGRID appliance Storage Node is the same whether you are recovering from the loss of the system drive or from the loss of storage volumes only.

As shown in the workflow diagram, you must prepare the appliance and reinstall software, configure the node to rejoin the grid, reformat storage, and restore object data.



Cautions and notes for recovering appliance Storage Nodes

If more than one Storage Node has failed (or is offline), contact technical support. Don't perform the following recovery procedure. Data loss could occur.

If this is the second Storage Node failure in less than 15 days after a Storage Node failure or recovery, contact technical support. Rebuilding Cassandra on two or more Storage Nodes within 15 days can result in data loss.

If more than one Storage Node at a site has failed, a site recovery procedure might be required. See How site recovery is performed by technical support.

If ILM rules are configured to store only one replicated copy and the copy exists on a storage volume that has failed, you will not be able to recover the object.

If you encounter a Services: Status - Cassandra (SVST) alarm during recovery, see Recover failed storage volumes and rebuild Cassandra database. After Cassandra is rebuilt, alarms should clear. If alarms don't clear, contact technical support.

For hardware maintenance procedures, such as instructions for replacing a controller or reinstalling SANtricity OS, see the maintenance instructions for your storage appliance.

- SGF6112 storage appliances
- SG6000 storage appliances
- SG5700 storage appliances

Prepare appliance Storage Node for reinstallation

When recovering an appliance Storage Node, you must first prepare the appliance for reinstallation of StorageGRID software.

Steps

1. Log in to the failed Storage Node:
   a. Enter the following command: `ssh admin@grid_node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

   When you are logged in as root, the prompt changes from `$` to `#`.

2. Prepare the appliance Storage Node for the installation of StorageGRID software. `sgareinstall`
3. When prompted to continue, enter: `y`

   The appliance reboots, and your SSH session ends. It usually takes about 5 minutes for the StorageGRID Appliance Installer to become available, although in some cases you might need to wait up to 30 minutes.

   Don't attempt to accelerate the reboot by cycling power or otherwise resetting the appliance. You might interrupt automatic BIOS, BMC, or other firmware upgrades.

   The StorageGRID appliance Storage Node is reset, and data on the Storage Node is no longer accessible. IP addresses configured during the original installation process should remain intact; however, it is recommended that you confirm this when the procedure completes.

   After executing the `sgareinstall` command, all StorageGRID-provisioned accounts, passwords, and SSH keys are removed, and new host keys are generated.

Start StorageGRID appliance installation

To install StorageGRID on an appliance Storage Node, you use the StorageGRID Appliance Installer, which is included on the appliance.

Before you begin

- The appliance has been installed in a rack, connected to your networks, and powered on.
- Network links and IP addresses have been configured for the appliance using the StorageGRID Appliance Installer.
- You know the IP address of the primary Admin Node for the StorageGRID grid.
- All Grid Network subnets listed on the IP Configuration page of the StorageGRID Appliance Installer have been defined in the Grid Network Subnet List on the primary Admin Node.
- You have completed these prerequisite tasks by following the installation instructions for your storage appliance. See Quick start for hardware installation.
- You are using a supported web browser.
- You know one of the IP addresses assigned to the compute controller in the appliance. You can use the IP address for the Admin Network (management port 1 on the controller), the Grid Network, or the Client Network.

About this task

To install StorageGRID on an appliance Storage Node:

- You specify or confirm the IP address of the primary Admin Node and the hostname (system name) of the node.
- You start the installation and wait as volumes are configured and the software is installed.
- Partway through the process, the installation pauses. To resume the installation, you must sign into the Grid Manager and configure the pending Storage Node as a replacement for the failed node.
- After you have configured the node, the appliance installation process completes, and the appliance is rebooted.

Steps

1. Open a browser and enter one of the IP addresses for the compute controller in the appliance.

   `https://Controller_IP:8443`

   The StorageGRID Appliance Installer Home page appears.

2. In the Primary Admin Node connection section, determine whether you need to specify the IP address for the primary Admin Node.

   The StorageGRID Appliance Installer can discover this IP address automatically, assuming the primary Admin Node, or at least one other grid node with ADMIN_IP configured, is present on the same subnet.

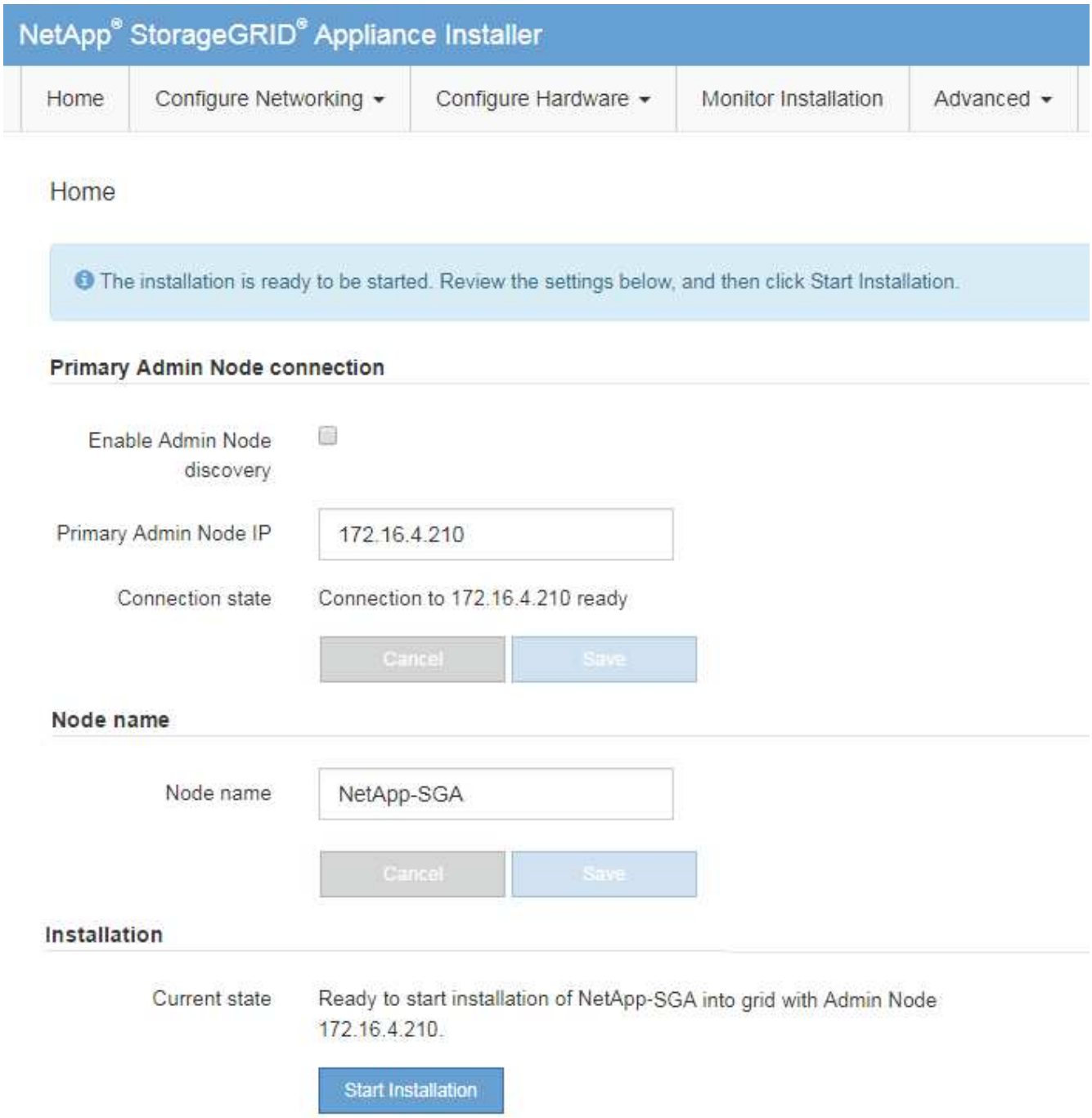
3. If this IP address is not shown or you need to change it, specify the address:

Option	Steps
Manual IP entry	a. Clear the **Enable Admin Node discovery** checkbox. b. Enter the IP address manually. c. Click **Save**. d. Wait while the connection state for the new IP address becomes “ready.”
Automatic discovery of all connected primary Admin Nodes	a. Select the **Enable Admin Node discovery** checkbox. b. From the list of discovered IP addresses, select the primary Admin Node for the grid where this appliance Storage Node will be deployed. c. Click **Save**. d. Wait while the connection state for the new IP address becomes “ready.”

4. In the **Node Name** field, enter the same hostname (system name) that was used for the node you are recovering, and click **Save**.
5. In the Installation section, confirm that the current state is “Ready to start installation of *node name* into grid with Primary Admin Node *admin_ip*” and that the **Start Installation** button is enabled.

   If the **Start Installation** button is not enabled, you might need to change the network configuration or port settings. For instructions, see the maintenance instructions for your appliance.

6. From the StorageGRID Appliance Installer home page, click **Start Installation**.



The Current state changes to “Installation is in progress,” and the Monitor Installation page is displayed.

If you need to access the Monitor Installation page manually, click **Monitor Installation** from the menu bar. See Monitor appliance installation.

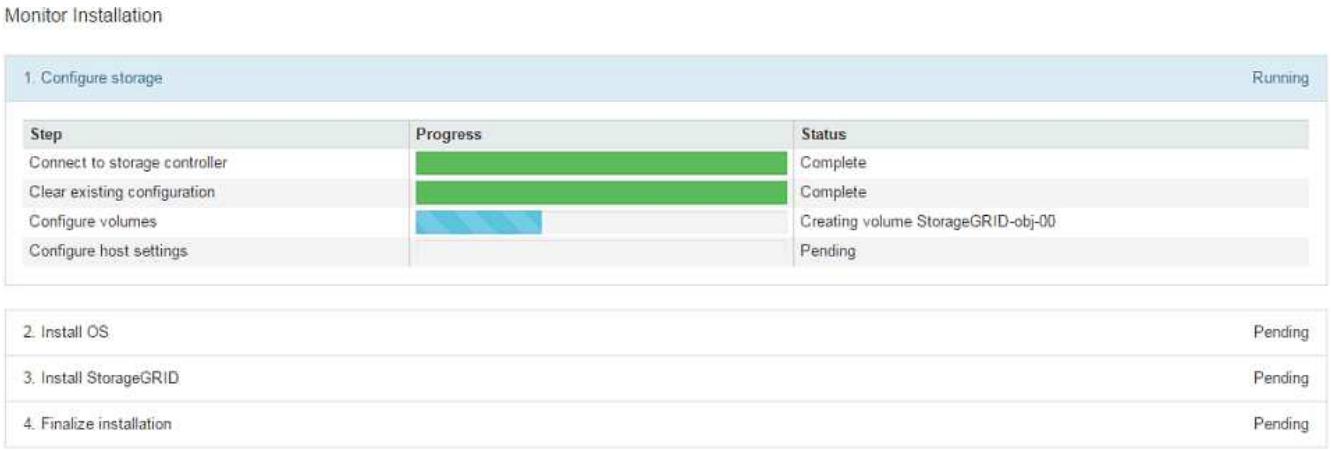
Monitor StorageGRID appliance installation

The StorageGRID Appliance Installer provides status until installation is complete. When the software installation is complete, the appliance is rebooted.

Steps

1. To monitor the installation progress, click **Monitor Installation** from the menu bar.

   The Monitor Installation page shows the installation progress.



   The blue status bar indicates which task is currently in progress. Green status bars indicate tasks that have completed successfully.

   > The installer ensures that tasks completed in a previous install aren't re-run. If you are re-running an installation, any tasks that don't need to be re-run are shown with a green status bar and a status of "Skipped."

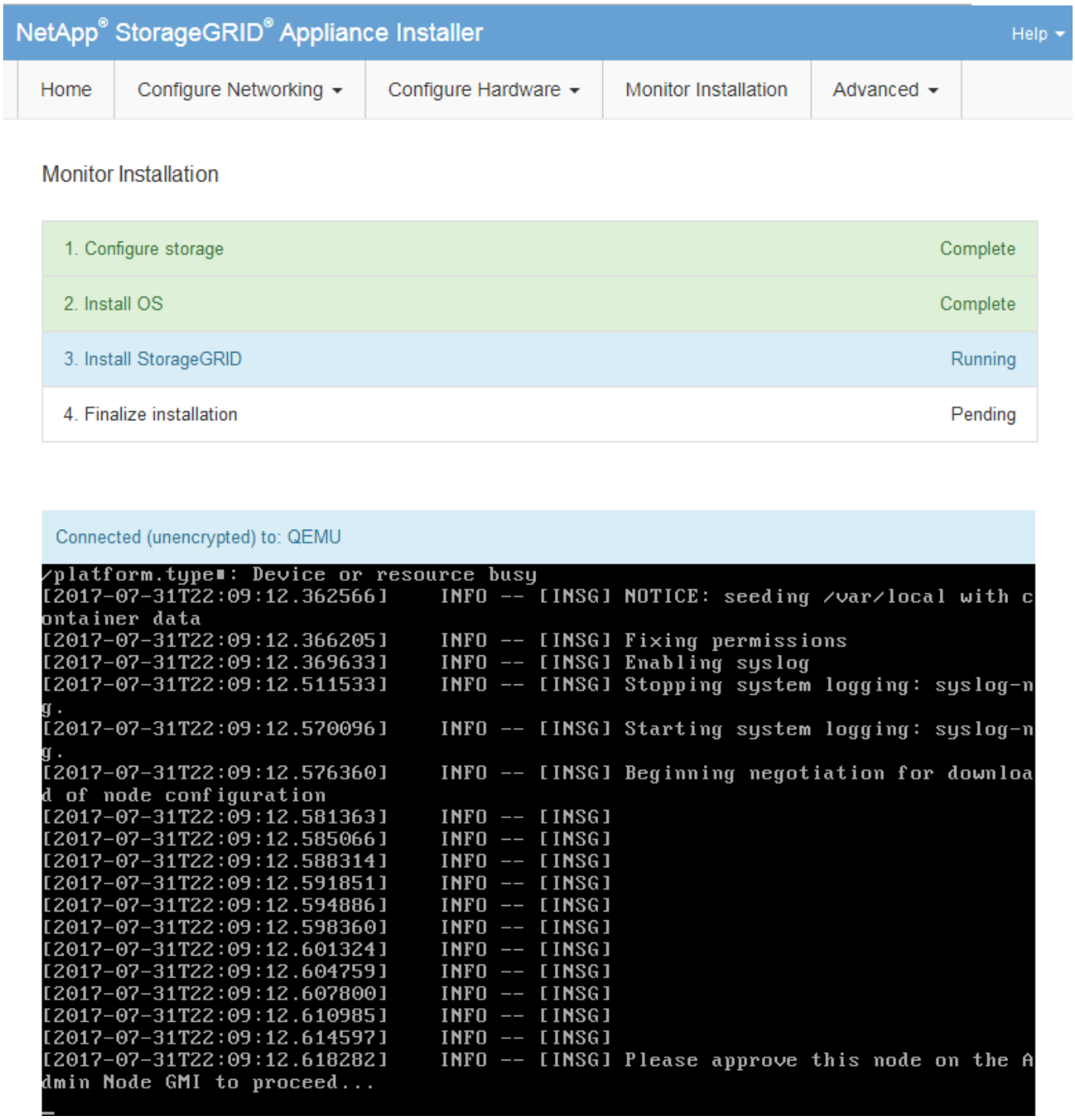
2. Review the progress of first two installation stages.
   - **1. Configure storage**

     During this stage, the installer connects to the storage controller, clears any existing configuration, communicates with SANtricity OS to configure volumes, and configures host settings.

   - **2. Install OS**

     During this stage, the installer copies the base operating system image for StorageGRID to the appliance.

3. Continue monitoring the installation progress until the **Install StorageGRID** stage pauses and a message appears on the embedded console prompting you to approve this node on the Admin Node using the Grid Manager.



4. Go to Select Start Recovery to configure appliance Storage Node.

Select Start Recovery to configure appliance Storage Node

You must select Start Recovery in the Grid Manager to configure an appliance Storage Node as a replacement for the failed node.

Before you begin

- You are signed in to the Grid Manager using a supported web browser.
- You have the Maintenance or Root access permission.
- You have the provisioning passphrase.

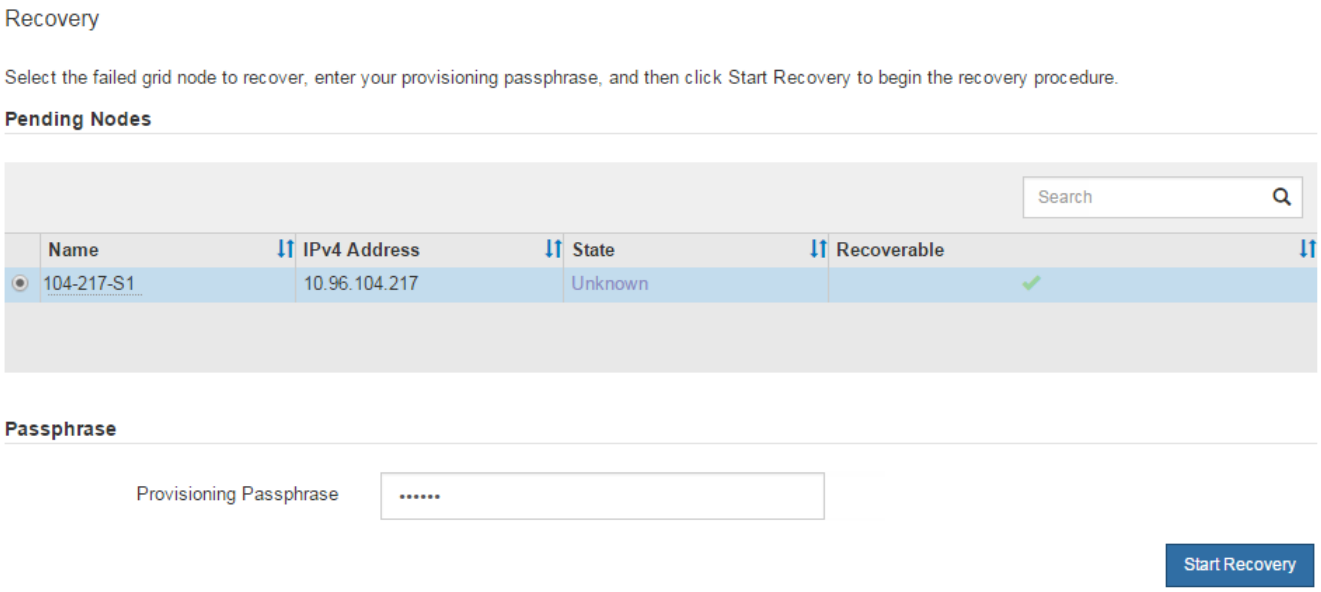
- You have deployed a recovery appliance Storage Node.
- You have the start date of any repair jobs for erasure-coded data.
- You have verified that the Storage Node has not been rebuilt within the last 15 days.

Steps

1. From the Grid Manager, select **MAINTENANCE** > **Tasks** > **Recovery**.
2. Select the grid node you want to recover in the Pending Nodes list.

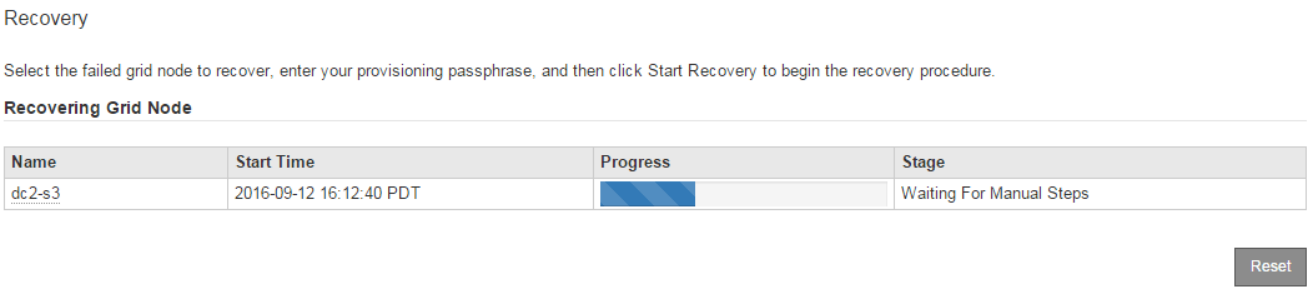
   Nodes appear in the list after they fail, but you can't select a node until it has been reinstalled and is ready for recovery.

3. Enter the **Provisioning Passphrase**.
4. Click **Start Recovery**.

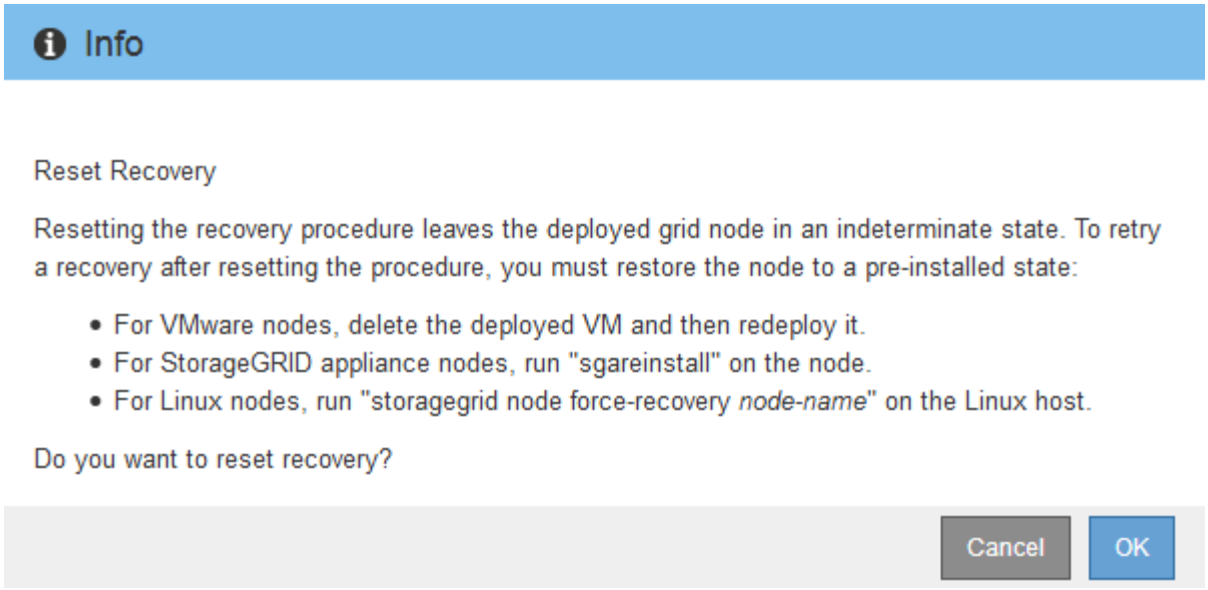


5. Monitor the progress of the recovery in the Recovering Grid Node table.

   When the grid node reaches the "Waiting for Manual Steps" stage, go to the next topic and perform the manual steps to remount and reformat appliance storage volumes.



At any point during the recovery, you can click **Reset** to start a new recovery. A dialog box appears, indicating that the node will be left in an indeterminate state if you reset the procedure.



If you want to retry the recovery after resetting the procedure, you must restore the appliance node to a pre-installed state by running `sgareinstall` on the node.

Remount and reformat appliance storage volumes (manual steps)

You must manually run two scripts to remount preserved storage volumes and reformat any failed storage volumes. The first script remounts volumes that are properly formatted as StorageGRID storage volumes. The second script reformats any unmounted volumes, rebuilds the Cassandra database, if needed, and starts services.

Before you begin

- You have already replaced the hardware for any failed storage volumes that you know require replacement.

  Running the `sn-remount-volumes` script might help you identify additional failed storage volumes.

- You have checked that a Storage Node decommissioning is not in progress, or you have paused the node decommission procedure. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Decommission**.)
- You have checked that an expansion is not in progress. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Expansion**.)

> Contact technical support if more than one Storage Node is offline or if a Storage Node in this grid has been rebuilt in the last 15 days. Don't run the `sn-recovery-postinstall.sh` script. Rebuilding Cassandra on two or more Storage Nodes within 15 days of each other might result in data loss.

About this task

To complete this procedure, you perform these high-level tasks:

- Log in to the recovered Storage Node.
- Run the `sn-remount-volumes` script to remount properly formatted storage volumes. When this script runs, it does the following:

- Mounts and unmounts each storage volume to replay the XFS journal.
- Performs an XFS file consistency check.
- If the file system is consistent, determines if the storage volume is a properly formatted StorageGRID storage volume.
- If the storage volume is properly formatted, remounts the storage volume. Any existing data on the volume remains intact.

- Review the script output and resolve any issues.
- Run the `sn-recovery-postinstall.sh` script. When this script runs, it does the following.

> ⓘ Don't reboot a Storage Node during recovery before running `sn-recovery-postinstall.sh` (step 4) to reformat the failed storage volumes and restore object metadata. Rebooting the Storage Node before `sn-recovery-postinstall.sh` completes causes errors for services that attempt to start and causes StorageGRID appliance nodes to exit maintenance mode.

  - Reformats any storage volumes that the `sn-remount-volumes` script could not mount or that were found to be improperly formatted.

    > ⓘ If a storage volume is reformatted, any data on that volume is lost. You must perform an additional procedure to restore object data from other locations in the grid, assuming that ILM rules were configured to store more than one object copy.

  - Rebuilds the Cassandra database on the node, if needed.
  - Starts the services on the Storage Node.

Steps

1. Log in to the recovered Storage Node:
   a. Enter the following command: `ssh admin@grid_node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

   When you are logged in as root, the prompt changes from `$` to `#`.

2. Run the first script to remount any properly formatted storage volumes.

   > ⓘ If all storage volumes are new and need to be formatted, or if all storage volumes have failed, you can skip this step and run the second script to reformat all unmounted storage volumes.

   a. Run the script: `sn-remount-volumes`

      This script might take hours to run on storage volumes that contain data.

   b. As the script runs, review the output and answer any prompts.

As required, you can use the `tail -f` command to monitor the contents of the script's log file (`/var/local/log/sn-remount-volumes.log`) . The log file contains more detailed information than the command line output.

```
root@SG:~ # sn-remount-volumes
The configured LDR noid is 12632740

====== Device /dev/sdb ======
Mount and unmount device /dev/sdb and checking file system
consistency:
The device is consistent.
Check rangedb structure on device /dev/sdb:
Mount device /dev/sdb to /tmp/sdb-654321 with rangedb mount options
This device has all rangedb directories.
Found LDR node id 12632740, volume number 0 in the volID file
Attempting to remount /dev/sdb
Device /dev/sdb remounted successfully

====== Device /dev/sdc ======
Mount and unmount device /dev/sdc and checking file system
consistency:
Error: File system consistency check retry failed on device /dev/sdc.
You can see the diagnosis information in the /var/local/log/sn-
remount-volumes.log.

This volume could be new or damaged. If you run sn-recovery-
postinstall.sh, this volume and any data on this volume will be
deleted. If you only had two copies of object data, you will
temporarily have only a single copy.
StorageGRID Webscale will attempt to restore data redundancy by
making additional replicated copies or EC fragments, according to the
rules in the active ILM policy.

Don't continue to the next step if you believe that the data
remaining on this volume can't be rebuilt from elsewhere in the grid
(for example, if your ILM policy uses a rule that makes only one copy
or if volumes have failed on multiple nodes). Instead, contact
support to determine how to recover your data.

====== Device /dev/sdd ======
Mount and unmount device /dev/sdd and checking file system
consistency:
Failed to mount device /dev/sdd
This device could be an uninitialized disk or has corrupted
superblock.
File system check might take a long time. Do you want to continue? (y
```

```
or n) [y/N]? y

Error: File system consistency check retry failed on device /dev/sdd.
You can see the diagnosis information in the /var/local/log/sn-
remount-volumes.log.

This volume could be new or damaged. If you run sn-recovery-
postinstall.sh, this volume and any data on this volume will be
deleted. If you only had two copies of object data, you will
temporarily have only a single copy.
StorageGRID Webscale will attempt to restore data redundancy by
making additional replicated copies or EC fragments, according to the
rules in the active ILM policy.

Don't continue to the next step if you believe that the data
remaining on this volume can't be rebuilt from elsewhere in the grid
(for example, if your ILM policy uses a rule that makes only one copy
or if volumes have failed on multiple nodes). Instead, contact
support to determine how to recover your data.

====== Device /dev/sde ======
Mount and unmount device /dev/sde and checking file system
consistency:
The device is consistent.
Check rangedb structure on device /dev/sde:
Mount device /dev/sde to /tmp/sde-654321 with rangedb mount options
This device has all rangedb directories.
Found LDR node id 12000078, volume number 9 in the volID file
Error: This volume does not belong to this node. Fix the attached
volume and re-run this script.
```

In the example output, one storage volume was remounted successfully and three storage volumes had errors.

- `/dev/sdb` passed the XFS file system consistency check and had a valid volume structure, so it was remounted successfully. Data on devices that are remounted by the script is preserved.
- `/dev/sdc` failed the XFS file system consistency check because the storage volume was new or corrupt.
- `/dev/sdd` could not be mounted because the disk was not initialized or the disk's superblock was corrupted. When the script can't mount a storage volume, it asks if you want to run the file system consistency check.
  - If the storage volume is attached to a new disk, answer **N** to the prompt. You don't need check the file system on a new disk.
  - If the storage volume is attached to an existing disk, answer **Y** to the prompt. You can use the results of the file system check to determine the source of the corruption. The results are saved in the `/var/local/log/sn-remount-volumes.log` log file.

- `/dev/sde` passed the XFS file system consistency check and had a valid volume structure; however, the LDR node ID in the `volID` file did not match the ID for this Storage Node (the `configured LDR noid` displayed at the top). This message indicates that this volume belongs to another Storage Node.

3. Review the script output and resolve any issues.

> If a storage volume failed the XFS file system consistency check or could not be mounted, carefully review the error messages in the output. You must understand the implications of running the `sn-recovery-postinstall.sh` script on these volumes.

a. Check to make sure that the results include an entry for all of the volumes you expected. If any volumes aren't listed, rerun the script.

b. Review the messages for all mounted devices. Make sure there are no errors indicating that a storage volume does not belong to this Storage Node.

In the example, the output for /dev/sde includes the following error message:

```
Error: This volume does not belong to this node. Fix the attached
volume and re-run this script.
```

> If a storage volume is reported as belonging to another Storage Node, contact technical support. If you run the `sn-recovery-postinstall.sh` script, the storage volume will be reformatted, which might cause data loss.

c. If any storage devices could not be mounted, make a note of the device name, and repair or replace the device.

> You must repair or replace any storage devices that could not be mounted.

You will use the device name to look up the volume ID, which is required input when you run the `repair-data` script to restore object data to the volume (the next procedure).

d. After repairing or replacing all unmountable devices, run the `sn-remount-volumes` script again to confirm that all storage volumes that can be remounted have been remounted.

> If a storage volume can't be mounted or is improperly formatted, and you continue to the next step, the volume and any data on the volume will be deleted. If you had two copies of object data, you will have only a single copy until you complete the next procedure (restoring object data).

> Don't run the `sn-recovery-postinstall.sh` script if you believe that the data remaining on a failed storage volume can't be rebuilt from elsewhere in the grid (for example, if your ILM policy uses a rule that makes only one copy or if volumes have failed on multiple nodes). Instead, contact technical support to determine how to recover your data.

4. Run the `sn-recovery-postinstall.sh` script: `sn-recovery-postinstall.sh`

This script reformats any storage volumes that could not be mounted or that were found to be improperly

formatted; rebuilds the Cassandra database on the node, if needed; and starts the services on the Storage Node.

Be aware of the following:

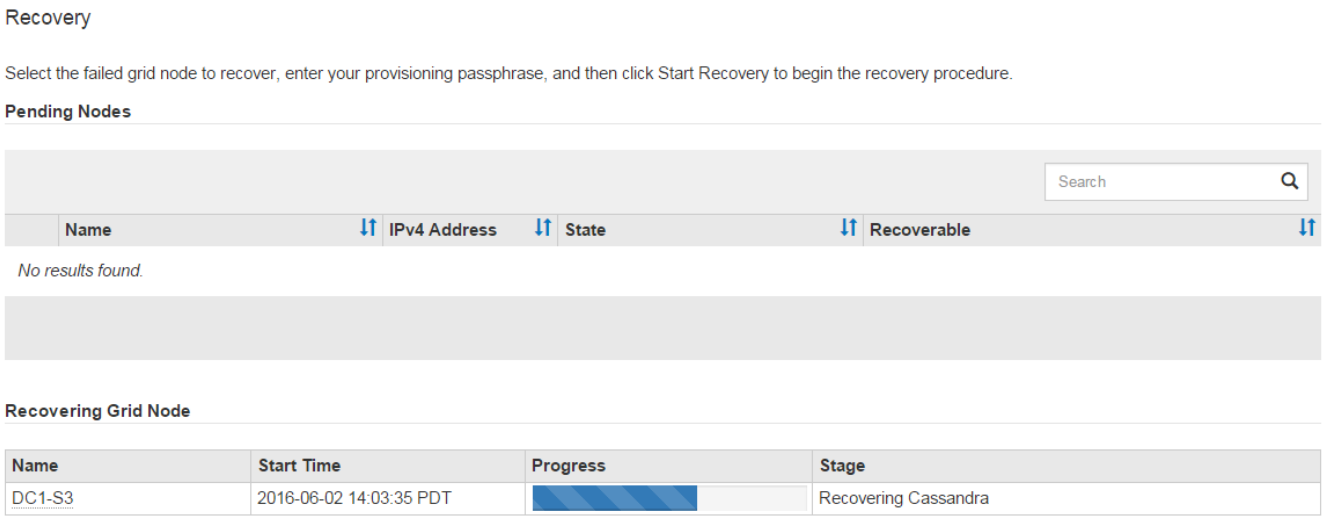
- The script might take hours to run.
- In general, you should leave the SSH session alone while the script is running.
- Don't press **Ctrl+C** while the SSH session is active.
- The script will run in the background if a network disruption occurs and terminates the SSH session, but you can view the progress from the Recovery page.
- If the Storage Node uses the RSM service, the script might appear to stall for 5 minutes as node services are restarted. This 5-minute delay is expected whenever the RSM service boots for the first time.

> The RSM service is present on Storage Nodes that include the ADC service.

> Some StorageGRID recovery procedures use Reaper to handle Cassandra repairs. Repairs occur automatically as soon as the related or required services have started. You might notice script output that mentions "reaper" or "Cassandra repair." If you see an error message indicating the repair has failed, run the command indicated in the error message.

5. As the `sn-recovery-postinstall.sh` script runs, monitor the Recovery page in the Grid Manager.

   The Progress bar and the Stage column on the Recovery page provide a high-level status of the `sn-recovery-postinstall.sh` script.



6. After the `sn-recovery-postinstall.sh` script has started services on the node, you can restore object data to any storage volumes that were formatted by the script.

   The script asks if you want to restore object data manually.

   - In most cases, you should restore object data using Grid Manager. Answer `n` to use the Grid Manager.
   - In rare cases, such as when instructed by technical support, or when you know that the replacement node has fewer volumes available for object storage than the original node, you must restore object data manually using the `repair-data` script. If one of these cases applies, answer `y`.

If you answer `y` to restore object data manually:

- You aren't able to restore object data using Grid Manager.
- You can monitor the progress of manual restoration jobs using Grid Manager.

Restore object data to storage volume for appliance

After recovering storage volumes for the appliance Storage Node, you can restore the replicated or erasure-coded object data that was lost when the Storage Node failed.

Which procedure should I use?

Whenever possible, restore object data using the **Volume restoration** page in the Grid Manager.

- If the volumes are listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, restore object data using the Volume restoration page in the Grid Manager.
- If the volumes aren't listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, follow the steps below for using the `repair-data` script to restore object data.

  If the recovered Storage Node contains fewer volumes than the node it is replacing, you must use the `repair-data` script.

Use the `repair-data` script to restore object data

Before you begin

- You must have confirmed that the recovered Storage Node has a Connection State of **Connected** on the **NODES** > **Overview** tab in the Grid Manager.

About this task

Object data can be restored from other Storage Nodes, an Archive Node, or a Cloud Storage Pool, assuming that the grid's ILM rules were configured such that object copies are available.

Note the following:

- If an ILM rule was configured to store only one replicated copy and that copy existed on a storage volume that failed, you will not be able to recover the object.
- If the only remaining copy of an object is in a Cloud Storage Pool, StorageGRID must issue multiple requests to the Cloud Storage Pool endpoint to restore object data. Before performing this procedure, contact technical support for help in estimating the recovery time frame and the associated costs.
- If the only remaining copy of an object is on an Archive Node, object data is retrieved from the Archive Node. Restoring object data to a Storage Node from an Archive Node takes longer than restoring copies from other Storage Nodes because of the latency associated with retrievals from external archival storage systems.

About the `repair-data` script

To restore object data, you run the `repair-data` script. This script begins the process of restoring object data and works with ILM scanning to ensure that ILM rules are met.

Select **Replicated data** or **Erasure-coded (EC) data** below to learn the different options for the `repair-data` script, based on whether you are restoring replicated data or erasure-coded data. If you need to restore both types of data, you must run both sets of commands.

For more information about the `repair-data` script, enter `repair-data --help` from the command line of the primary Admin Node.

Replicated data

Two commands are available for restoring replicated data, based on whether you need to repair the entire node or only certain volumes on the node:

```
repair-data start-replicated-node-repair

repair-data start-replicated-volume-repair
```

You can track repairs of replicated data with this command:

```
repair-data show-replicated-repair-status
```

Erasure coded (EC) data

Two commands are available for restoring erasure-coded data, based on whether you need to repair the entire node or only certain volumes on the node:

```
repair-data start-ec-node-repair

repair-data start-ec-volume-repair
```

You can track repairs of erasure-coded data with this command:

```
repair-data show-ec-repair-status
```

Repairs of erasure-coded data can begin while some Storage Nodes are offline. However, if all erasure-coded data can't be accounted for, the repair can't be completed. Repair will complete after all nodes are available.

The EC repair job temporarily reserves a large amount of storage. Storage alerts might be triggered, but will resolve when the repair is complete. If there is not enough storage for the reservation, the EC repair job will fail. Storage reservations are released when the EC repair job completes, whether the job failed or succeeded.

Find hostname for Storage Node

1. Log in to the primary Admin Node:
   a. Enter the following command: `ssh admin@primary_Admin_Node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

When you are logged in as root, the prompt changes from `$` to `#`.

2. Use the `/etc/hosts` file to find the hostname of the Storage Node for the restored storage volumes. To see a list of all nodes in the grid, enter the following: `cat /etc/hosts`.

Repair data if all volumes have failed

If all storage volumes have failed, repair the entire node. Follow the instructions for **replicated data**, **erasure-coded (EC) data**, or both, based on whether you use replicated data, erasure-coded (EC) data, or both.

If only some volumes have failed, go to Repair data if only some volumes have failed.

> You can't run `repair-data` operations for more than one node at the same time. To recover multiple nodes, contact technical support.

Replicated data

If your grid includes replicated data, use the `repair-data start-replicated-node-repair` command with the `--nodes` option, where `--nodes` is the hostname (system name), to repair the entire Storage Node.

This command repairs the replicated data on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-node-repair --nodes SG-DC-SN3
```

> As object data is restored, the **Objects Lost** alert is triggered if the StorageGRID system can't locate replicated object data. Alerts might be triggered on Storage Nodes throughout the system. You should determine the cause of the loss and if recovery is possible. See Investigate lost objects.

Erasure coded (EC) data

If your grid contains erasure-coded data, use the `repair-data start-ec-node-repair` command with the `--nodes` option, where `--nodes` is the hostname (system name), to repair the entire Storage Node.

This command repairs the erasure-coded data on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-node-repair --nodes SG-DC-SN3
```

The operation returns a unique `repair ID` that identifies this `repair_data` operation. Use this `repair ID` to track the progress and result of the `repair_data` operation. No other feedback is returned as the recovery process completes.

> Repairs of erasure-coded data can begin while some Storage Nodes are offline. Repair will complete after all nodes are available.

Repair data if only some volumes have failed

If only some of the volumes have failed, repair the affected volumes. Follow the instructions for **replicated data**, **erasure-coded (EC) data**, or both, based on whether you use replicated data, erasure-coded (EC) data, or both.

If all volumes have failed, go to Repair data if all volumes have failed.

Enter the volume IDs in hexadecimal. For example, `0000` is the first volume and `000F` is the sixteenth volume. You can specify one volume, a range of volumes, or multiple volumes that aren't in a sequence.

All the volumes must be on the same Storage Node. If you need to restore volumes for more than one Storage Node, contact technical support.

Replicated data

If your grid contains replicated data, use the `start-replicated-volume-repair` command with the `--nodes` option to identify the node (where `--nodes` is the hostname of the node). Then add either the `--volumes` or `--volume-range` option, as shown in the following examples.

Single volume: This command restores replicated data to volume `0002` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volumes 0002
```

Range of volumes: This command restores replicated data to all volumes in the range `0003` to `0009` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volume-range
0003,0009
```

Multiple volumes not in a sequence: This command restores replicated data to volumes `0001`, `0005`, and `0008` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volumes
0001,0005,0008
```

> As object data is restored, the **Objects Lost** alert is triggered if the StorageGRID system can't locate replicated object data. Alerts might be triggered on Storage Nodes throughout the system. Note the alert description and recommended actions to determine the cause of the loss and if recovery is possible.

Erasure coded (EC) data

If your grid contains erasure-coded data, use the `start-ec-volume-repair` command with the `--nodes` option to identify the node (where `--nodes` is the hostname of the node). Then add either the `--volumes` or `--volume-range` option, as shown in the following examples.

Single volume: This command restores erasure-coded data to volume `0007` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volumes 0007
```

Range of volumes: This command restores erasure-coded data to all volumes in the range `0004` to `0006` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volume-range 0004,0006
```

Multiple volumes not in a sequence: This command restores erasure-coded data to volumes `000A`, `000C`, and `000E` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volumes 000A,000C,000E
```

The `repair-data` operation returns a unique `repair ID` that identifies this `repair_data` operation. Use this `repair ID` to track the progress and result of the `repair_data` operation. No other feedback is returned as the recovery process completes.

Repairs of erasure-coded data can begin while some Storage Nodes are offline. Repair will complete after all nodes are available.

Monitor repairs

Monitor the status of the repair jobs, based on whether you use **replicated data**, **erasure-coded (EC) data**, or both.

You can also monitor the status of volume restoration jobs in process and view a history of restoration jobs completed in Grid Manager.

Replicated data

- To get an estimated percent completion for the replicated repair, add the `show-replicated-repair-status` option to the repair-data command.

  ```
  repair-data show-replicated-repair-status
  ```

- To determine if repairs are complete:
  1. Select **NODES** > ***Storage Node being repaired*** > **ILM**.
  2. Review the attributes in the Evaluation section. When repairs are complete, the **Awaiting - All** attribute indicates 0 objects.
- To monitor the repair in more detail:
  1. Select **SUPPORT** > **Tools** > **Grid topology**.
  2. Select ***grid*** > ***Storage Node being repaired*** > **LDR** > **Data Store**.
  3. Use a combination of the following attributes to determine, as well as possible, if replicated repairs are complete.

     Cassandra inconsistencies might be present, and failed repairs aren't tracked.

     - **Repairs Attempted (XRPA)**: Use this attribute to track the progress of replicated repairs. This attribute increases each time a Storage Node tries to repair a high-risk object. When this attribute does not increase for a period longer than the current scan period (provided by the **Scan Period — Estimated** attribute), it means that ILM scanning found no high-risk objects that need to be repaired on any nodes.

       High-risk objects are objects that are at risk of being completely lost. This does not include objects that don't satisfy their ILM configuration.

     - **Scan Period — Estimated (XSCM)**: Use this attribute to estimate when a policy change will be applied to previously ingested objects. If the **Repairs Attempted** attribute does not increase for a period longer than the current scan period, it is probable that replicated repairs are done. Note that the scan period can change. The **Scan Period — Estimated (XSCM)** attribute applies to the entire grid and is the maximum of all node scan periods. You can query the **Scan Period — Estimated** attribute history for the grid to determine an appropriate time frame.

Erasure coded (EC) data

To monitor the repair of erasure-coded data and retry any requests that might have failed:

1. Determine the status of erasure-coded data repairs:
   - Select **SUPPORT** > **Tools** > **Metrics** to view the estimated time to completion and the completion percentage for the current job. Then, select **EC Overview** in the Grafana section. Look at the **Grid EC Job Estimated Time to Completion** and **Grid EC Job Percentage Completed** dashboards.
   - Use this command to see the status of a specific `repair-data` operation:

     ```
     repair-data show-ec-repair-status --repair-id repair ID
     ```

   - Use this command to list all repairs:

```
repair-data show-ec-repair-status
```

The output lists information, including `repair ID`, for all previously and currently running repairs.

2. If the output shows that the repair operation failed, use the `--repair-id` option to retry the repair.

   This command retries a failed node repair, using the repair ID 6949309319275667690:

   ```
   repair-data start-ec-node-repair --repair-id 6949309319275667690
   ```

   This command retries a failed volume repair, using the repair ID 6949309319275667690:

   ```
   repair-data start-ec-volume-repair --repair-id 6949309319275667690
   ```

Check storage state after recovering appliance Storage Node

After recovering an appliance Storage Node, you must verify that the desired state of the appliance Storage Node is set to online and ensure that the state will be online by default whenever the Storage Node server is restarted.

Before you begin

- You are signed in to the Grid Manager using a supported web browser.
- The Storage Node has been recovered, and data recovery is complete.

Steps

1. Select **SUPPORT** > **Tools** > **Grid topology**.
2. Check the values of **Recovered Storage Node** > **LDR** > **Storage** > **Storage State — Desired** and **Storage State — Current**.

   The value of both attributes should be Online.

3. If the Storage State — Desired is set to Read-only, complete the following steps:
   a. Click the **Configuration** tab.
   b. From the **Storage State — Desired** drop-down list, select **Online**.
   c. Click **Apply Changes**.
   d. Click the **Overview** tab and confirm that the values of **Storage State — Desired** and **Storage State — Current** are updated to Online.

Recover from storage volume failure where system drive is intact

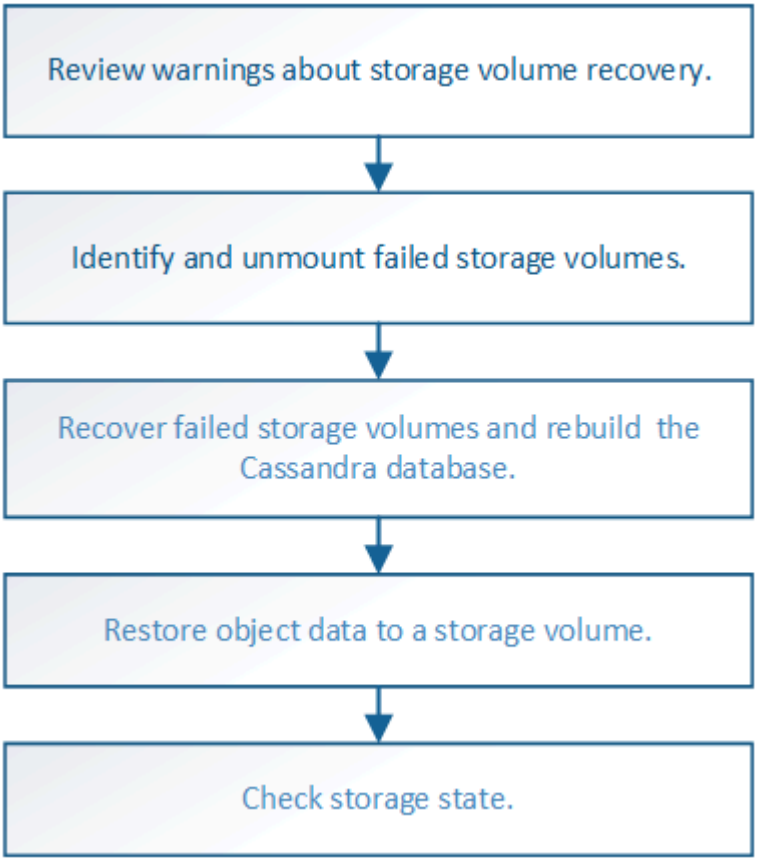
Recover from storage volume failure where system drive is intact: Workflow

You must complete a series of tasks to recover a software-based Storage Node where one or more storage volumes on the Storage Node have failed, but the system drive is intact. If only storage volumes have failed, the Storage Node is still available to the

StorageGRID system.

> This recovery procedure applies to software-based Storage Nodes only. If storage volumes have failed on an appliance Storage Node, use the appliance procedure instead: Recover appliance Storage Node.

As shown in the workflow diagram, you must identify and unmount failed storage volumes, recover the volumes, rebuild the Cassandra database, and restore object data.



Review warnings about storage volume recovery

Before recovering failed storage volumes for a Storage Node, you must review the following warnings.

The storage volumes (or rangedbs) in a Storage Node are identified by a hexadecimal number, which is known as the volume ID. For example, 0000 is the first volume and 000F is the sixteenth volume. The first object store (volume 0) on each Storage Node uses up to 4 TB of space for object metadata and Cassandra database operations; any remaining space on that volume is used for object data. All other storage volumes are used exclusively for object data.

If volume 0 fails and needs to be recovered, the Cassandra database might be rebuilt as part of the volume recovery procedure. Cassandra might also be rebuilt in the following circumstances:

- A Storage Node is brought back online after having been offline for more than 15 days.
- The system drive and one or more storage volumes fails and is recovered.

When Cassandra is rebuilt, the system uses information from other Storage Nodes. If too many Storage Nodes

are offline, some Cassandra data might not be available. If Cassandra has been rebuilt recently, Cassandra data might not yet be consistent across the grid. Data loss can occur if Cassandra is rebuilt when too many Storage Nodes are offline or if two or more Storage Nodes are rebuilt within 15 days of each other.

> If more than one Storage Node has failed (or is offline), contact technical support. Don't perform the following recovery procedure. Data loss could occur.

> If this is the second Storage Node failure in less than 15 days after a Storage Node failure or recovery, contact technical support. Rebuilding Cassandra on two or more Storage Nodes within 15 days can result in data loss.

> If more than one Storage Node at a site has failed, a site recovery procedure might be required. See How site recovery is performed by technical support.

> If ILM rules are configured to store only one replicated copy and the copy exists on a storage volume that has failed, you will not be able to recover the object.

> If you encounter a Services: Status - Cassandra (SVST) alarm during recovery, see Recover failed storage volumes and rebuild Cassandra database. After Cassandra is rebuilt, alarms should clear. If alarms don't clear, contact technical support.

Related information

Warnings and considerations for grid node recovery

Identify and unmount failed storage volumes

When recovering a Storage Node with failed storage volumes, you must identify and unmount the failed volumes. You must verify that only the failed storage volumes are reformatted as part of the recovery procedure.

Before you begin

You are signed in to the Grid Manager using a supported web browser.

About this task

You should recover failed storage volumes as soon as possible.

The first step of the recovery process is to detect volumes that have become detached, need to be unmounted, or have I/O errors. If failed volumes are still attached but have a randomly corrupted file system, the system might not detect any corruption in unused or unallocated parts of the disk.

> You must finish this procedure before performing manual steps to recover the volumes, such as adding or re-attaching the disks, stopping the node, starting the node, or rebooting. Otherwise, when you run the `reformat_storage_block_devices.rb` script, you might encounter a file system error that causes the script to hang or fail.

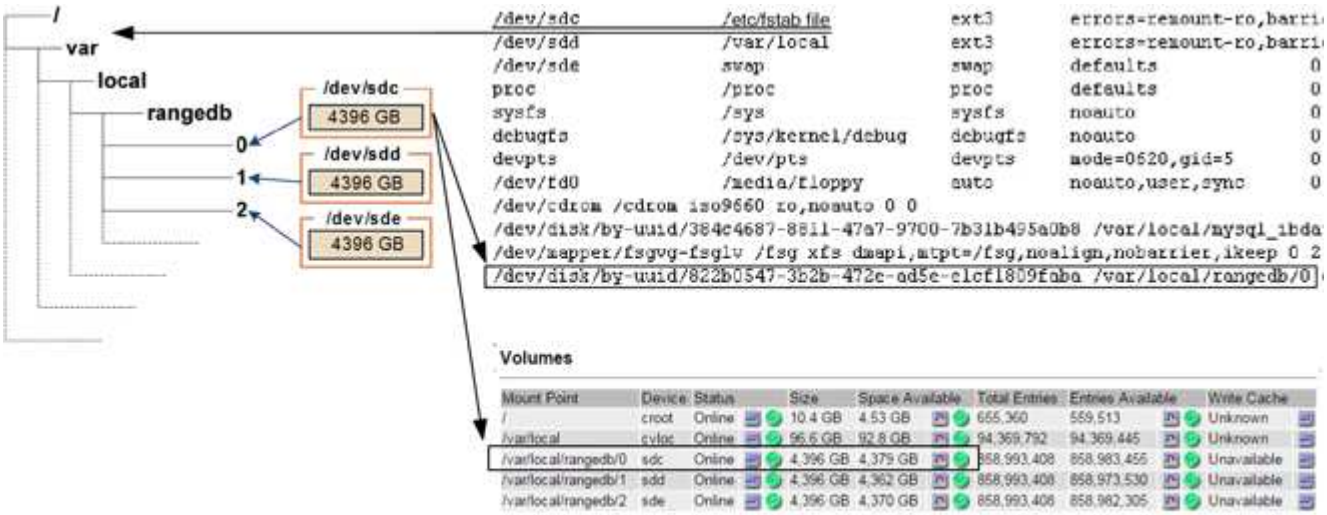
> Repair the hardware and properly attach the disks before running the `reboot` command.

Identify failed storage volumes carefully. You will use this information to verify which volumes must be reformatted. Once a volume has been reformatted, data on the volume can't be recovered.

To correctly recover failed storage volumes, you need to know both the device names of the failed storage volumes and their volume IDs.

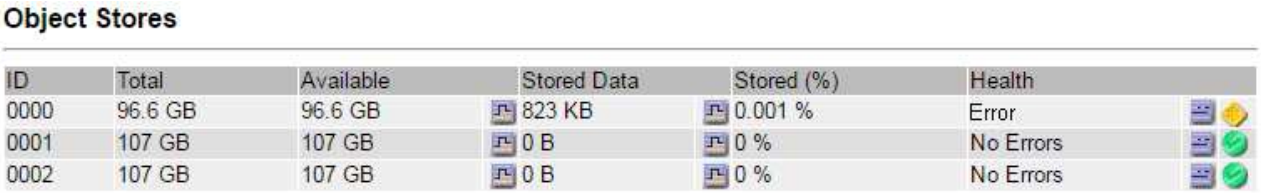
At installation, each storage device is assigned a file system universal unique identifier (UUID) and is mounted to a rangedb directory on the Storage Node using that assigned file system UUID. The file system UUID and the rangedb directory are listed in the `/etc/fstab` file. The device name, rangedb directory, and the size of the mounted volume are displayed in the Grid Manager.

In the following example, device `/dev/sdc` has a volume size of 4 TB, is mounted to `/var/local/rangedb/0`, using the device name `/dev/disk/by-uuid/822b0547-3b2b-472e-ad5e-e1cf1809faba` in the `/etc/fstab` file:



Steps

1. Complete the following steps to record the failed storage volumes and their device names:
   a. Select **SUPPORT** > **Tools** > **Grid topology**.
   b. Select **site** > **failed Storage Node** > **LDR** > **Storage** > **Overview** > **Main**, and look for object stores with alarms.



   c. Select **site** > **failed Storage Node** > **SSM** > **Resources** > **Overview** > **Main**. Determine the mount point and volume size of each failed storage volume identified in the previous step.

      Object stores are numbered in hex notation. For example, 0000 is the first volume and 000F is the sixteenth volume. In the example, the object store with an ID of 0000 corresponds to

`/var/local/rangedb/0` with device name sdc and a size of 107 GB.

Volumes

Mount Point	Device	Status	Size	Space Available	Total Entries	Entries Available	Write Cache
/	croot	Online	10.4 GB	4.17 GB	655,360	554,806	Unknown
/var/local	cvloc	Online	96.6 GB	96.1 GB	94,369,792	94,369,423	Unknown
/var/local/rangedb/0	sdc	Online	107 GB	107 GB	104,857,600	104,856,202	Enabled
/var/local/rangedb/1	sdd	Online	107 GB	107 GB	104,857,600	104,856,536	Enabled
/var/local/rangedb/2	sde	Online	107 GB	107 GB	104,857,600	104,856,536	Enabled

2. Log in to the failed Storage Node:

   a. Enter the following command: `ssh admin@grid_node_IP`

   b. Enter the password listed in the `Passwords.txt` file.

   c. Enter the following command to switch to root: `su -`

   d. Enter the password listed in the `Passwords.txt` file.

   When you are logged in as root, the prompt changes from `$` to `#`.

3. Run the following script to unmount a failed storage volume:

   ```
   sn-unmount-volume object_store_ID
   ```

   The `object_store_ID` is the ID of the failed storage volume. For example, specify `0` in the command for an object store with ID 0000.

4. If prompted, press **y** to stop the Cassandra service depending on storage volume 0.

   If the Cassandra service is already stopped, you aren't prompted. The Cassandra service is stopped only for volume 0.

   ```
   root@Storage-180:~/var/local/tmp/storage~ # sn-unmount-volume 0
   Services depending on storage volume 0 (cassandra) aren't down.
   Services depending on storage volume 0 must be stopped before running
   this script.
   Stop services that require storage volume 0 [y/N]? y
   Shutting down services that require storage volume 0.
   Services requiring storage volume 0 stopped.
   Unmounting /var/local/rangedb/0
   /var/local/rangedb/0 is unmounted.
   ```

   In a few seconds, the volume is unmounted. Messages appear indicating each step of the process. The final message indicates that the volume is unmounted.

5. If the unmount fails because the volume is busy, you can force an unmount using the `--use-umountof` option:

Forcing an unmount using the `--use-umountof` option might cause processes or services using the volume to behave unexpectedly or crash.

```
root@Storage-180:~ # sn-unmount-volume --use-umountof
/var/local/rangedb/2
Unmounting /var/local/rangedb/2 using umountof
/var/local/rangedb/2 is unmounted.
Informing LDR service of changes to storage volumes
```

Recover failed storage volumes and rebuild Cassandra database

You must run a script that reformats and remounts storage on failed storage volumes, and rebuilds the Cassandra database on the Storage Node if the system determines that it is necessary.

Before you begin

- You have the `Passwords.txt` file.
- The system drives on the server are intact.
- The cause of the failure has been identified and, if necessary, replacement storage hardware has already been acquired.
- The total size of the replacement storage is the same as the original.
- You have checked that a Storage Node decommissioning is not in progress, or you have paused the node decommission procedure. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Decommission**.)
- You have checked that an expansion is not in progress. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Expansion**.)
- You have reviewed the warnings about storage volume recovery.

Steps

1. As needed, replace failed physical or virtual storage associated with the failed storage volumes that you identified and unmounted earlier.

   Don't remount the volumes in this step. The storage is remounted and added to `/etc/fstab` in a later step.

2. In the Grid Manager, go to **NODES** > **`appliance Storage Node`** > **Hardware**. In the StorageGRID Appliance section of the page, verify that the Storage RAID mode is healthy.
3. Log in to the failed Storage Node:
   a. Enter the following command: `ssh admin@grid_node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

      When you are logged in as root, the prompt changes from `$` to `#`.

4. Use a text editor (vi or vim) to delete failed volumes from the `/etc/fstab` file and then save the file.

> Commenting out a failed volume in the `/etc/fstab` file is insufficient. The volume must be deleted from `fstab` as the recovery process verifies that all lines in the `fstab` file match the mounted file systems.

5. Reformat any failed storage volumes and rebuild the Cassandra database if it is necessary. Enter: `reformat_storage_block_devices.rb`
   - When storage volume 0 is unmounted, prompts and messages will indicate that the Cassandra service is being stopped.
   - You will be prompted to rebuild the Cassandra database if it is necessary.
     - Review the warnings. If none of them apply, rebuild the Cassandra database. Enter: **y**
     - If more than one Storage Node is offline or if another Storage Node has been rebuilt in the last 15 days. Enter: **n**

       The script will exit without rebuilding Cassandra. Contact technical support.
   - For each rangedb drive on the Storage Node, when you are asked: `Reformat the rangedb drive <name> (device <major number>:<minor number>)? [y/n]?`, enter one of the following responses:
     - **y** to reformat a drive that had errors. This reformats the storage volume and adds the reformatted storage volume to the `/etc/fstab` file.
     - **n** if the drive contains no errors, and you don't want to reformat it.

       > Selecting **n** exits the script. Either mount the drive (if you think the data on the drive should be retained and the drive was unmounted in error) or remove the drive. Then, run the `reformat_storage_block_devices.rb` command again.

       > Some StorageGRID recovery procedures use Reaper to handle Cassandra repairs. Repairs occur automatically as soon as the related or required services have started. You might notice script output that mentions "reaper" or "Cassandra repair." If you see an error message indicating the repair has failed, run the command indicated in the error message.

     In the following example output, the drive `/dev/sdf` must be reformatted, and Cassandra did not need to be rebuilt:

```
root@DC1-S1:~ # reformat_storage_block_devices.rb
Formatting devices that are not in use...
Skipping in use device /dev/sdc
Skipping in use device /dev/sdd
Skipping in use device /dev/sde
Reformat the rangedb drive /dev/sdf (device 8:64)? [Y/n]? y
Successfully formatted /dev/sdf with UUID b951bfcb-4804-41ad-b490-
805dfd8df16c
All devices processed
Running: /usr/local/ldr/setup_rangedb.sh 12368435
Cassandra does not need rebuilding.
Starting services.
Informing storage services of new volume

Reformatting done.  Now do manual steps to
restore copies of data.
```

After the storage volumes are reformatted and remounted and necessary Cassandra operations are complete, you can restore object data using Grid Manager.

Restore object data to storage volume where system drive is intact

After recovering a storage volume on a Storage Node where the system drive is intact, you can restore the replicated or erasure-coded object data that was lost when the storage volume failed.

Which procedure should I use?

Whenever possible, restore object data using the **Volume restoration** page in the Grid Manager.

- If the volumes are listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, restore object data using the Volume restoration page in the Grid Manager.
- If the volumes aren't listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, follow the steps below for using the `repair-data` script to restore object data.

  If the recovered Storage Node contains fewer volumes than the node it is replacing, you must use the `repair-data` script.

Use the `repair-data` script to restore object data

Before you begin

- You must have confirmed that the recovered Storage Node has a Connection State of **Connected** on the **NODES** > **Overview** tab in the Grid Manager.

About this task

Object data can be restored from other Storage Nodes, an Archive Node, or a Cloud Storage Pool, assuming

that the grid's ILM rules were configured such that object copies are available.

Note the following:

- If an ILM rule was configured to store only one replicated copy and that copy existed on a storage volume that failed, you will not be able to recover the object.
- If the only remaining copy of an object is in a Cloud Storage Pool, StorageGRID must issue multiple requests to the Cloud Storage Pool endpoint to restore object data. Before performing this procedure, contact technical support for help in estimating the recovery time frame and the associated costs.
- If the only remaining copy of an object is on an Archive Node, object data is retrieved from the Archive Node. Restoring object data to a Storage Node from an Archive Node takes longer than restoring copies from other Storage Nodes because of the latency associated with retrievals from external archival storage systems.

About the `repair-data` script

To restore object data, you run the `repair-data` script. This script begins the process of restoring object data and works with ILM scanning to ensure that ILM rules are met.

Select **Replicated data** or **Erasure-coded (EC) data** below to learn the different options for the `repair-data` script, based on whether you are restoring replicated data or erasure-coded data. If you need to restore both types of data, you must run both sets of commands.

> For more information about the `repair-data` script, enter `repair-data --help` from the command line of the primary Admin Node.

Replicated data

Two commands are available for restoring replicated data, based on whether you need to repair the entire node or only certain volumes on the node:

```
repair-data start-replicated-node-repair

repair-data start-replicated-volume-repair
```

You can track repairs of replicated data with this command:

```
repair-data show-replicated-repair-status
```

Erasure coded (EC) data

Two commands are available for restoring erasure-coded data, based on whether you need to repair the entire node or only certain volumes on the node:

```
repair-data start-ec-node-repair

repair-data start-ec-volume-repair
```

You can track repairs of erasure-coded data with this command:

```
repair-data show-ec-repair-status
```

> Repairs of erasure-coded data can begin while some Storage Nodes are offline. However, if all erasure-coded data can't be accounted for, the repair can't be completed. Repair will complete after all nodes are available.

> The EC repair job temporarily reserves a large amount of storage. Storage alerts might be triggered, but will resolve when the repair is complete. If there is not enough storage for the reservation, the EC repair job will fail. Storage reservations are released when the EC repair job completes, whether the job failed or succeeded.

Find hostname for Storage Node

1. Log in to the primary Admin Node:
   a. Enter the following command: `ssh admin@primary_Admin_Node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

   When you are logged in as root, the prompt changes from `$` to `#`.

2. Use the `/etc/hosts` file to find the hostname of the Storage Node for the restored storage volumes. To see a list of all nodes in the grid, enter the following: `cat /etc/hosts`.

Repair data if all volumes have failed

If all storage volumes have failed, repair the entire node. Follow the instructions for **replicated data**, **erasure-coded (EC) data**, or both, based on whether you use replicated data, erasure-coded (EC) data, or both.

If only some volumes have failed, go to Repair data if only some volumes have failed.

> You can't run `repair-data` operations for more than one node at the same time. To recover multiple nodes, contact technical support.

Replicated data

If your grid includes replicated data, use the `repair-data start-replicated-node-repair` command with the `--nodes` option, where `--nodes` is the hostname (system name), to repair the entire Storage Node.

This command repairs the replicated data on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-node-repair --nodes SG-DC-SN3
```

> As object data is restored, the **Objects Lost** alert is triggered if the StorageGRID system can't locate replicated object data. Alerts might be triggered on Storage Nodes throughout the system. You should determine the cause of the loss and if recovery is possible. See Investigate lost objects.

Erasure coded (EC) data

If your grid contains erasure-coded data, use the `repair-data start-ec-node-repair` command with the `--nodes` option, where `--nodes` is the hostname (system name), to repair the entire Storage Node.

This command repairs the erasure-coded data on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-node-repair --nodes SG-DC-SN3
```

The operation returns a unique `repair ID` that identifies this `repair_data` operation. Use this `repair ID` to track the progress and result of the `repair_data` operation. No other feedback is returned as the recovery process completes.

> Repairs of erasure-coded data can begin while some Storage Nodes are offline. Repair will complete after all nodes are available.

Repair data if only some volumes have failed

If only some of the volumes have failed, repair the affected volumes. Follow the instructions for **replicated data**, **erasure-coded (EC) data**, or both, based on whether you use replicated data, erasure-coded (EC) data, or both.

If all volumes have failed, go to Repair data if all volumes have failed.

Enter the volume IDs in hexadecimal. For example, `0000` is the first volume and `000F` is the sixteenth volume. You can specify one volume, a range of volumes, or multiple volumes that aren't in a sequence.

All the volumes must be on the same Storage Node. If you need to restore volumes for more than one Storage Node, contact technical support.

Replicated data

If your grid contains replicated data, use the `start-replicated-volume-repair` command with the `--nodes` option to identify the node (where `--nodes` is the hostname of the node). Then add either the `--volumes` or `--volume-range` option, as shown in the following examples.

Single volume: This command restores replicated data to volume `0002` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volumes 0002
```

Range of volumes: This command restores replicated data to all volumes in the range `0003` to `0009` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volume-range
0003,0009
```

Multiple volumes not in a sequence: This command restores replicated data to volumes `0001`, `0005`, and `0008` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volumes
0001,0005,0008
```

> As object data is restored, the **Objects Lost** alert is triggered if the StorageGRID system can't locate replicated object data. Alerts might be triggered on Storage Nodes throughout the system. Note the alert description and recommended actions to determine the cause of the loss and if recovery is possible.

Erasure coded (EC) data

If your grid contains erasure-coded data, use the `start-ec-volume-repair` command with the `--nodes` option to identify the node (where `--nodes` is the hostname of the node). Then add either the `--volumes` or `--volume-range` option, as shown in the following examples.

Single volume: This command restores erasure-coded data to volume `0007` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volumes 0007
```

Range of volumes: This command restores erasure-coded data to all volumes in the range `0004` to `0006` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volume-range 0004,0006
```

Multiple volumes not in a sequence: This command restores erasure-coded data to volumes `000A`, `000C`, and `000E` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volumes 000A,000C,000E
```

The `repair-data` operation returns a unique `repair ID` that identifies this `repair_data` operation. Use this `repair ID` to track the progress and result of the `repair_data` operation. No other feedback is returned as the recovery process completes.

Repairs of erasure-coded data can begin while some Storage Nodes are offline. Repair will complete after all nodes are available.

Monitor repairs

Monitor the status of the repair jobs, based on whether you use **replicated data**, **erasure-coded (EC) data**, or both.

You can also monitor the status of volume restoration jobs in process and view a history of restoration jobs completed in Grid Manager.

Replicated data

- To get an estimated percent completion for the replicated repair, add the `show-replicated-repair-status` option to the repair-data command.

  ```
  repair-data show-replicated-repair-status
  ```

- To determine if repairs are complete:
  1. Select **NODES** > ***Storage Node being repaired*** > **ILM**.
  2. Review the attributes in the Evaluation section. When repairs are complete, the **Awaiting - All** attribute indicates 0 objects.
- To monitor the repair in more detail:
  1. Select **SUPPORT** > **Tools** > **Grid topology**.
  2. Select ***grid*** > ***Storage Node being repaired*** > **LDR** > **Data Store**.
  3. Use a combination of the following attributes to determine, as well as possible, if replicated repairs are complete.

     Cassandra inconsistencies might be present, and failed repairs aren't tracked.

     - **Repairs Attempted (XRPA)**: Use this attribute to track the progress of replicated repairs. This attribute increases each time a Storage Node tries to repair a high-risk object. When this attribute does not increase for a period longer than the current scan period (provided by the **Scan Period — Estimated** attribute), it means that ILM scanning found no high-risk objects that need to be repaired on any nodes.

       High-risk objects are objects that are at risk of being completely lost. This does not include objects that don't satisfy their ILM configuration.

     - **Scan Period — Estimated (XSCM)**: Use this attribute to estimate when a policy change will be applied to previously ingested objects. If the **Repairs Attempted** attribute does not increase for a period longer than the current scan period, it is probable that replicated repairs are done. Note that the scan period can change. The **Scan Period — Estimated (XSCM)** attribute applies to the entire grid and is the maximum of all node scan periods. You can query the **Scan Period — Estimated** attribute history for the grid to determine an appropriate time frame.

Erasure coded (EC) data

To monitor the repair of erasure-coded data and retry any requests that might have failed:

1. Determine the status of erasure-coded data repairs:
   - Select **SUPPORT** > **Tools** > **Metrics** to view the estimated time to completion and the completion percentage for the current job. Then, select **EC Overview** in the Grafana section. Look at the **Grid EC Job Estimated Time to Completion** and **Grid EC Job Percentage Completed** dashboards.
   - Use this command to see the status of a specific `repair-data` operation:

     ```
     repair-data show-ec-repair-status --repair-id repair ID
     ```

   - Use this command to list all repairs:

```
repair-data show-ec-repair-status
```

The output lists information, including `repair ID`, for all previously and currently running repairs.

2. If the output shows that the repair operation failed, use the `--repair-id` option to retry the repair.

   This command retries a failed node repair, using the repair ID 6949309319275667690:

   ```
   repair-data start-ec-node-repair --repair-id 6949309319275667690
   ```

   This command retries a failed volume repair, using the repair ID 6949309319275667690:

   ```
   repair-data start-ec-volume-repair --repair-id 6949309319275667690
   ```

Check storage state after recovering storage volumes

After recovering storage volumes, you must verify that the desired state of the Storage Node is set to online and ensure that the state will be online by default whenever the Storage Node server is restarted.

Before you begin

- You are signed in to the Grid Manager using a supported web browser.
- The Storage Node has been recovered, and data recovery is complete.

Steps

1. Select **SUPPORT** > **Tools** > **Grid topology**.
2. Check the values of **Recovered Storage Node** > **LDR** > **Storage** > **Storage State — Desired** and **Storage State — Current**.

   The value of both attributes should be Online.

3. If the Storage State — Desired is set to Read-only, complete the following steps:
   a. Click the **Configuration** tab.
   b. From the **Storage State — Desired** drop-down list, select **Online**.
   c. Click **Apply Changes**.
   d. Click the **Overview** tab and confirm that the values of **Storage State — Desired** and **Storage State — Current** are updated to Online.

Recover from system drive failure

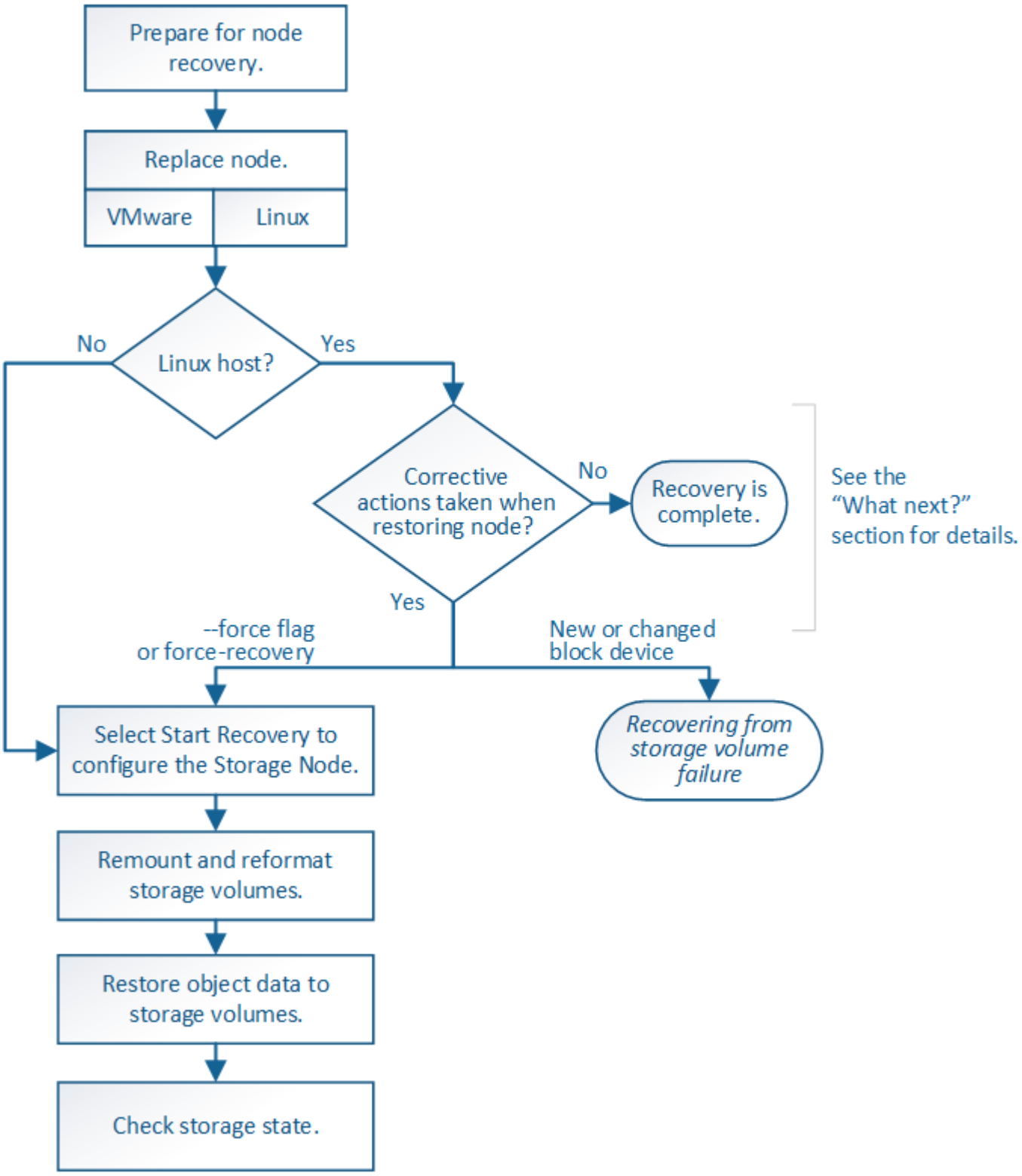
Recover from system drive failure: Workflow

If the system drive on a software-based Storage Node has failed, the Storage Node is not available to the StorageGRID system. You must complete a specific set of tasks to recover from a system drive failure.

Use this procedure to recover from a system drive failure on a software-based Storage Node. This procedure

includes the steps to follow if any storage volumes also failed or can't be remounted.

This procedure applies to software-based Storage Nodes only. You must follow a different procedure to recover an appliance Storage Node.



Review warnings for Storage Node system drive recovery

Before recovering a failed system drive of a Storage Node, review the general warnings and considerations for grid node recovery and the following specific warnings.

Storage Nodes have a Cassandra database that includes object metadata. The Cassandra database might be rebuilt in the following circumstances:

- A Storage Node is brought back online after having been offline for more than 15 days.
- A storage volume has failed and been recovered.
- The system drive and one or more storage volumes fails and is recovered.

When Cassandra is rebuilt, the system uses information from other Storage Nodes. If too many Storage Nodes are offline, some Cassandra data might not be available. If Cassandra has been rebuilt recently, Cassandra data might not yet be consistent across the grid. Data loss can occur if Cassandra is rebuilt when too many Storage Nodes are offline or if two or more Storage Nodes are rebuilt within 15 days of each other.

If more than one Storage Node has failed (or is offline), contact technical support. Don't perform the following recovery procedure. Data loss could occur.

If this is the second Storage Node failure in less than 15 days after a Storage Node failure or recovery, contact technical support. Rebuilding Cassandra on two or more Storage Nodes within 15 days can result in data loss.

If more than one Storage Node at a site has failed, a site recovery procedure might be required. See How site recovery is performed by technical support.

If this Storage Node is in read-only maintenance mode to allow for the retrieval of objects by another Storage Node with failed storage volumes, recover volumes on the Storage Node with failed storage volumes before recovering this failed Storage Node. See the instructions to recover from storage volume failure where system drive is intact.

If ILM rules are configured to store only one replicated copy and the copy exists on a storage volume that has failed, you will not be able to recover the object.

If you encounter a Services: Status - Cassandra (SVST) alarm during recovery, see Recover failed storage volumes and rebuild Cassandra database. After Cassandra is rebuilt, alarms should clear. If alarms don't clear, contact technical support.

Replace the Storage Node

If the system drive has failed, you must first replace the Storage Node.

You must select the node replacement procedure for your platform. The steps to replace a node are the same for all types of grid nodes.

This procedure applies to software-based Storage Nodes only. You must follow a different procedure to recover an appliance Storage Node.

Linux: If you aren't sure if your system drive has failed, follow the instructions to replace the node to determine which recovery steps are required.

Platform	Procedure
VMware	Replace a VMware node
Linux	Replace a Linux node
OpenStack	NetApp-provided virtual machine disk files and scripts for OpenStack are no longer supported for recovery operations. If you need to recover a node running in an OpenStack deployment, download the files for your Linux operating system. Then, follow the procedure for replacing a Linux node.

Select Start Recovery to configure Storage Node

After replacing a Storage Node, you must select Start Recovery in the Grid Manager to configure the new node as a replacement for the failed node.

Before you begin

- You are signed in to the Grid Manager using a supported web browser.
- You have the Maintenance or Root access permission.
- You have the provisioning passphrase.
- You have deployed and configured the replacement node.
- You have the start date of any repair jobs for erasure-coded data.
- You have verified that the Storage Node has not been rebuilt within the last 15 days.

About this task

If the Storage Node is installed as a container on a Linux host, you must perform this step only if one of these is true:

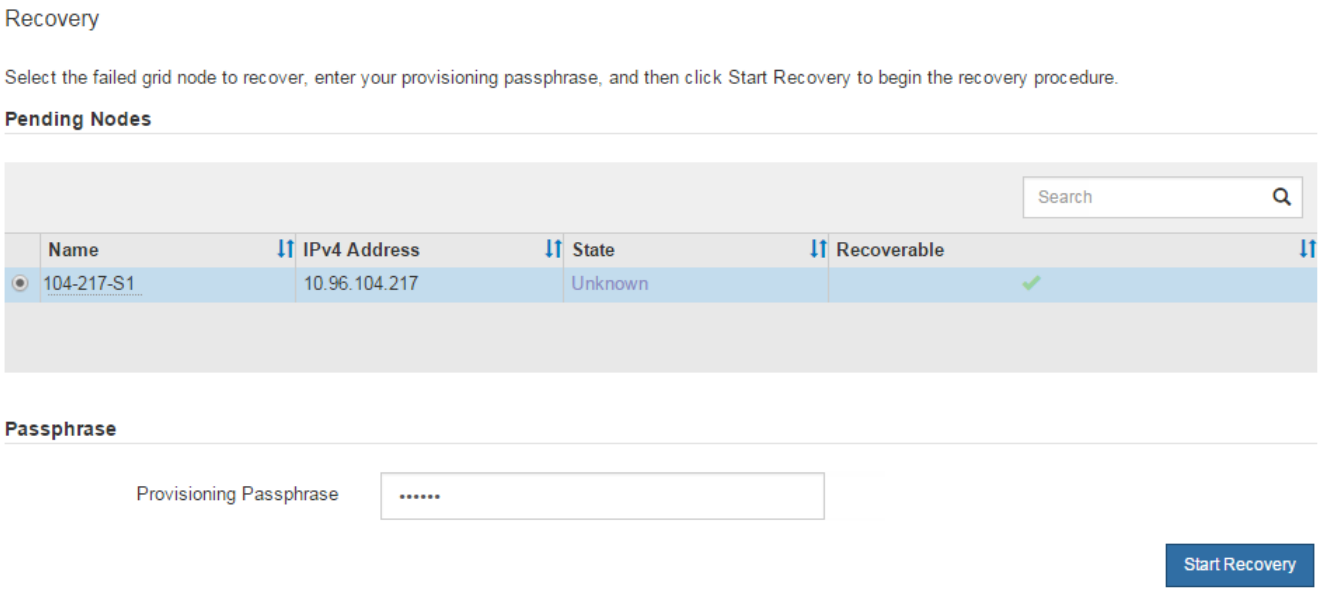
- You had to use the `--force` flag to import the node, or you issued `storagegrid node force-recovery node-name`
- You had to do a full node reinstall, or you needed to restore /var/local.

Steps

1. From the Grid Manager, select **MAINTENANCE** > **Tasks** > **Recovery**.
2. Select the grid node you want to recover in the Pending Nodes list.

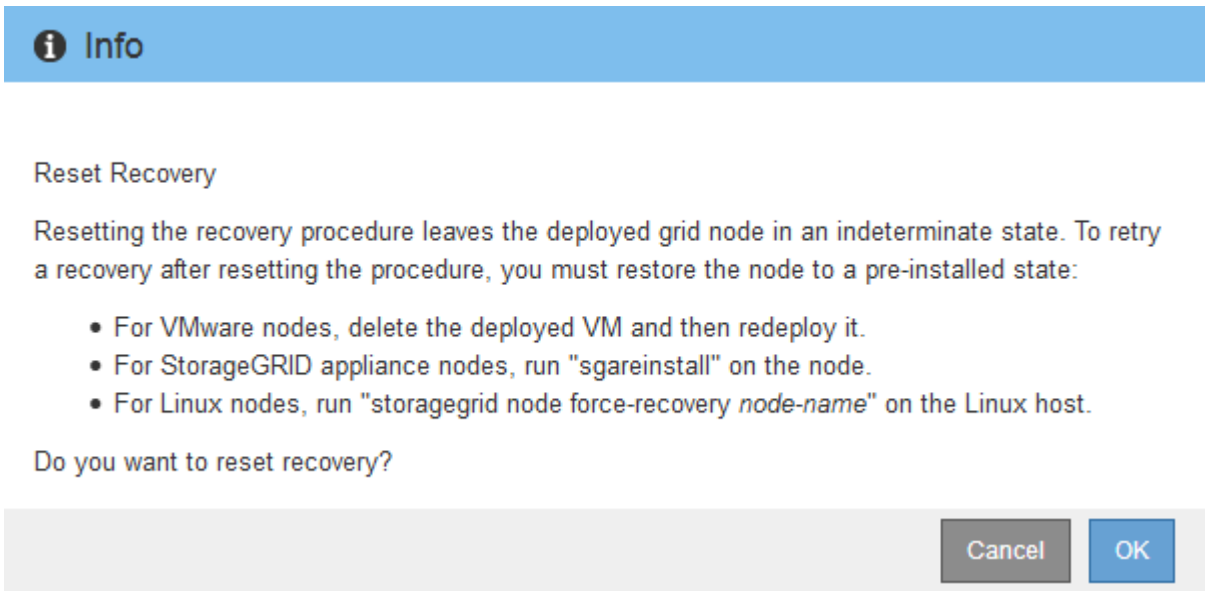
   Nodes appear in the list after they fail, but you can't select a node until it has been reinstalled and is ready for recovery.

3. Enter the **Provisioning Passphrase**.
4. Click **Start Recovery**.



5. Monitor the progress of the recovery in the Recovering Grid Node table.

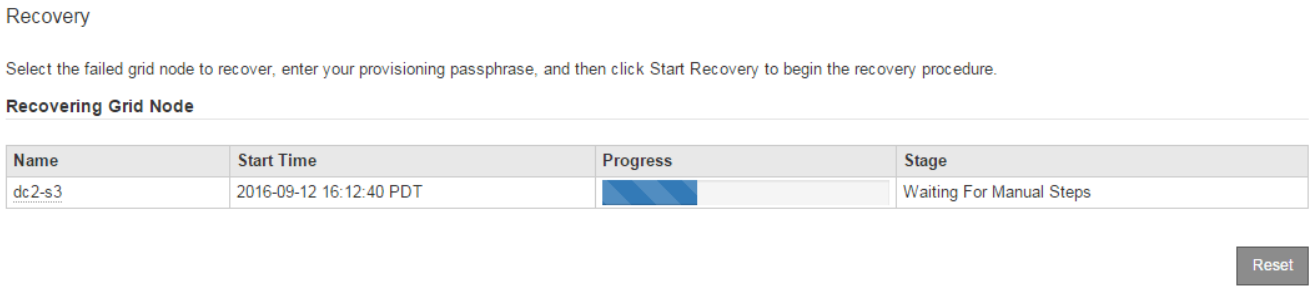
> While the recovery procedure is running, you can click **Reset** to start a new recovery. A dialog box appears, indicating that the node will be left in an indeterminate state if you reset the procedure.



If you want to retry the recovery after resetting the procedure, you must restore the node to a pre-installed state, as follows:

- **VMware**: Delete the deployed virtual grid node. Then, when you are ready to restart the recovery, redeploy the node.
- **Linux**: Restart the node by running this command on the Linux host: `storagegrid node force-recovery node-name`

6. When the Storage Node reaches the “Waiting for Manual Steps” stage, go to Remount and reformat storage volumes (manual steps).



Remount and reformat storage volumes (manual steps)

You must manually run two scripts to remount preserved storage volumes and to reformat any failed storage volumes. The first script remounts volumes that are properly formatted as StorageGRID storage volumes. The second script reformats any unmounted volumes, rebuilds Cassandra, if needed, and starts services.

Before you begin

- You have already replaced the hardware for any failed storage volumes that you know require replacement.

  Running the `sn-remount-volumes` script might help you identify additional failed storage volumes.

- You have checked that a Storage Node decommissioning is not in progress, or you have paused the node decommission procedure. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Decommission**.)
- You have checked that an expansion is not in progress. (In the Grid Manager, select **MAINTENANCE** > **Tasks** > **Expansion**.)
- You have reviewed the warnings for Storage Node system drive recovery.

> Contact technical support if more than one Storage Node is offline or if a Storage Node in this grid has been rebuilt in the last 15 days. Don't run the `sn-recovery-postinstall.sh` script. Rebuilding Cassandra on two or more Storage Nodes within 15 days of each other might result in data loss.

About this task

To complete this procedure, you perform these high-level tasks:

- Log in to the recovered Storage Node.
- Run the `sn-remount-volumes` script to remount properly formatted storage volumes. When this script runs, it does the following:
  - Mounts and unmounts each storage volume to replay the XFS journal.
  - Performs an XFS file consistency check.
  - If the file system is consistent, determines if the storage volume is a properly formatted StorageGRID storage volume.
  - If the storage volume is properly formatted, remounts the storage volume. Any existing data on the volume remains intact.
- Review the script output and resolve any issues.

- Run the `sn-recovery-postinstall.sh` script. When this script runs, it does the following.

> Don't reboot a Storage Node during recovery before running `sn-recovery-postinstall.sh` to reformat the failed storage volumes and restore object metadata. Rebooting the Storage Node before `sn-recovery-postinstall.sh` completes causes errors for services that attempt to start and causes StorageGRID appliance nodes to exit maintenance mode. See the step for post-install script.

  - Reformats any storage volumes that the `sn-remount-volumes` script could not mount or that were found to be improperly formatted.

> If a storage volume is reformatted, any data on that volume is lost. You must perform an additional procedure to restore object data from other locations in the grid, assuming that ILM rules were configured to store more than one object copy.

  - Rebuilds the Cassandra database on the node, if needed.
  - Starts the services on the Storage Node.

Steps

1. Log in to the recovered Storage Node:
   a. Enter the following command: `ssh admin@grid_node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

   When you are logged in as root, the prompt changes from `$` to `#`.

2. Run the first script to remount any properly formatted storage volumes.

> If all storage volumes are new and need to be formatted, or if all storage volumes have failed, you can skip this step and run the second script to reformat all unmounted storage volumes.

   a. Run the script: `sn-remount-volumes`

      This script might take hours to run on storage volumes that contain data.

   b. As the script runs, review the output and answer any prompts.

> As required, you can use the `tail -f` command to monitor the contents of the script's log file (`/var/local/log/sn-remount-volumes.log`) . The log file contains more detailed information than the command line output.

```
root@SG:~ # sn-remount-volumes
The configured LDR noid is 12632740

====== Device /dev/sdb ======
```

```
Mount and unmount device /dev/sdb and checking file system
consistency:
The device is consistent.
Check rangedb structure on device /dev/sdb:
Mount device /dev/sdb to /tmp/sdb-654321 with rangedb mount options
This device has all rangedb directories.
Found LDR node id 12632740, volume number 0 in the volID file
Attempting to remount /dev/sdb
Device /dev/sdb remounted successfully

====== Device /dev/sdc ======
Mount and unmount device /dev/sdc and checking file system
consistency:
Error: File system consistency check retry failed on device /dev/sdc.
You can see the diagnosis information in the /var/local/log/sn-
remount-volumes.log.

This volume could be new or damaged. If you run sn-recovery-
postinstall.sh,
this volume and any data on this volume will be deleted. If you only
had two
copies of object data, you will temporarily have only a single copy.
StorageGRID Webscale will attempt to restore data redundancy by
making
additional replicated copies or EC fragments, according to the rules
in
the active ILM policy.

Don't continue to the next step if you believe that the data
remaining on
this volume can't be rebuilt from elsewhere in the grid (for example,
if
your ILM policy uses a rule that makes only one copy or if volumes
have
failed on multiple nodes). Instead, contact support to determine how
to
recover your data.

====== Device /dev/sdd ======
Mount and unmount device /dev/sdd and checking file system
consistency:
Failed to mount device /dev/sdd
This device could be an uninitialized disk or has corrupted
superblock.
File system check might take a long time. Do you want to continue? (y
or n) [y/N]? y
```

```
Error: File system consistency check retry failed on device /dev/sdd.
You can see the diagnosis information in the /var/local/log/sn-
remount-volumes.log.

This volume could be new or damaged. If you run sn-recovery-
postinstall.sh,
this volume and any data on this volume will be deleted. If you only
had two
copies of object data, you will temporarily have only a single copy.
StorageGRID Webscale will attempt to restore data redundancy by
making
additional replicated copies or EC fragments, according to the rules
in
the active ILM policy.

Don't continue to the next step if you believe that the data
remaining on
this volume can't be rebuilt from elsewhere in the grid (for example,
if
your ILM policy uses a rule that makes only one copy or if volumes
have
failed on multiple nodes). Instead, contact support to determine how
to
recover your data.

====== Device /dev/sde ======
Mount and unmount device /dev/sde and checking file system
consistency:
The device is consistent.
Check rangedb structure on device /dev/sde:
Mount device /dev/sde to /tmp/sde-654321 with rangedb mount options
This device has all rangedb directories.
Found LDR node id 12000078, volume number 9 in the volID file
Error: This volume does not belong to this node. Fix the attached
volume and re-run this script.
```

In the example output, one storage volume was remounted successfully and three storage volumes had errors.

- `/dev/sdb` passed the XFS file system consistency check and had a valid volume structure, so it was remounted successfully. Data on devices that are remounted by the script is preserved.
- `/dev/sdc` failed the XFS file system consistency check because the storage volume was new or corrupt.
- `/dev/sdd` could not be mounted because the disk was not initialized or the disk's superblock was corrupted. When the script can't mount a storage volume, it asks if you want to run the file system consistency check.

- If the storage volume is attached to a new disk, answer **N** to the prompt. You don't need check the file system on a new disk.
- If the storage volume is attached to an existing disk, answer **Y** to the prompt. You can use the results of the file system check to determine the source of the corruption. The results are saved in the `/var/local/log/sn-remount-volumes.log` log file.

- `/dev/sde` passed the XFS file system consistency check and had a valid volume structure; however, the LDR node ID in the volID file did not match the ID for this Storage Node (the `configured LDR noid` displayed at the top). This message indicates that this volume belongs to another Storage Node.

3. Review the script output and resolve any issues.

> If a storage volume failed the XFS file system consistency check or could not be mounted, carefully review the error messages in the output. You must understand the implications of running the `sn-recovery-postinstall.sh` script on these volumes.

a. Check to make sure that the results include an entry for all of the volumes you expected. If any volumes aren't listed, rerun the script.

b. Review the messages for all mounted devices. Make sure there are no errors indicating that a storage volume does not belong to this Storage Node.

   In the example, the output for `/dev/sde` includes the following error message:

```
Error: This volume does not belong to this node. Fix the attached
volume and re-run this script.
```

> If a storage volume is reported as belonging to another Storage Node, contact technical support. If you run the `sn-recovery-postinstall.sh` script, the storage volume will be reformatted, which might cause data loss.

c. If any storage devices could not be mounted, make a note of the device name, and repair or replace the device.

> You must repair or replace any storage devices that could not be mounted.

   You will use the device name to look up the volume ID, which is required input when you run the `repair-data` script to restore object data to the volume (the next procedure).

d. After repairing or replacing all unmountable devices, run the `sn-remount-volumes` script again to confirm that all storage volumes that can be remounted have been remounted.

> If a storage volume can't be mounted or is improperly formatted, and you continue to the next step, the volume and any data on the volume will be deleted. If you had two copies of object data, you will have only a single copy until you complete the next procedure (restoring object data).

Don't run the `sn-recovery-postinstall.sh` script if you believe that the data remaining on a failed storage volume can't be rebuilt from elsewhere in the grid (for example, if your ILM policy uses a rule that makes only one copy or if volumes have failed on multiple nodes). Instead, contact technical support to determine how to recover your data.

4. Run the `sn-recovery-postinstall.sh` script: `sn-recovery-postinstall.sh`

   This script reformats any storage volumes that could not be mounted or that were found to be improperly formatted; rebuilds the Cassandra database on the node, if needed; and starts the services on the Storage Node.

   Be aware of the following:

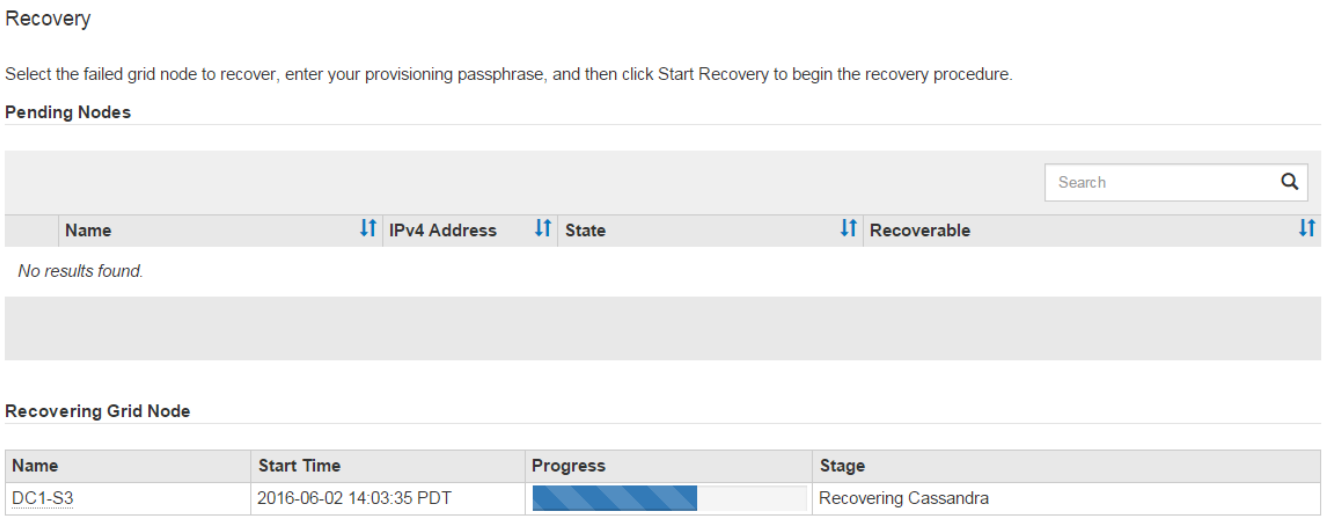
   - The script might take hours to run.
   - In general, you should leave the SSH session alone while the script is running.
   - Don't press **Ctrl+C** while the SSH session is active.
   - The script will run in the background if a network disruption occurs and terminates the SSH session, but you can view the progress from the Recovery page.
   - If the Storage Node uses the RSM service, the script might appear to stall for 5 minutes as node services are restarted. This 5-minute delay is expected whenever the RSM service boots for the first time.

     The RSM service is present on Storage Nodes that include the ADC service.

   Some StorageGRID recovery procedures use Reaper to handle Cassandra repairs. Repairs occur automatically as soon as the related or required services have started. You might notice script output that mentions "reaper" or "Cassandra repair." If you see an error message indicating the repair has failed, run the command indicated in the error message.

5. As the `sn-recovery-postinstall.sh` script runs, monitor the Recovery page in the Grid Manager.

   The Progress bar and the Stage column on the Recovery page provide a high-level status of the `sn-recovery-postinstall.sh` script.



6. After the `sn-recovery-postinstall.sh` script has started services on the node, you can restore object data to any storage volumes that were formatted by the script.

   The script asks if you want to restore object data manually.

   - In most cases, you should restore object data using Grid Manager. Answer `n` to use the Grid Manager.
   - In rare cases, such as when instructed by technical support, or when you know that the replacement node has fewer volumes available for object storage than the original node, you must restore object data manually using the `repair-data` script. If one of these cases applies, answer `y`.

   > If you answer `y` to restore object data manually:
   >
   > - You aren't able to restore object data using Grid Manager.
   > - You can monitor the progress of manual restoration jobs using Grid Manager.

Restore object data to storage volume (system drive failure)

After recovering storage volumes for a non-appliance Storage Node, you can restore the replicated or erasure-coded object data that was lost when the Storage Node failed.

Which procedure should I use?

Whenever possible, restore object data using the **Volume restoration** page in the Grid Manager.

- If the volumes are listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, restore object data using the Volume restoration page in the Grid Manager.
- If the volumes aren't listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, follow the steps below for using the `repair-data` script to restore object data.

  If the recovered Storage Node contains fewer volumes than the node it is replacing, you must use the `repair-data` script.

Use the `repair-data` script to restore object data

Before you begin

- You must have confirmed that the recovered Storage Node has a Connection State of **Connected** on the **NODES** > **Overview** tab in the Grid Manager.

About this task

Object data can be restored from other Storage Nodes, an Archive Node, or a Cloud Storage Pool, assuming that the grid's ILM rules were configured such that object copies are available.

Note the following:

- If an ILM rule was configured to store only one replicated copy and that copy existed on a storage volume that failed, you will not be able to recover the object.
- If the only remaining copy of an object is in a Cloud Storage Pool, StorageGRID must issue multiple requests to the Cloud Storage Pool endpoint to restore object data. Before performing this procedure, contact technical support for help in estimating the recovery time frame and the associated costs.

- If the only remaining copy of an object is on an Archive Node, object data is retrieved from the Archive Node. Restoring object data to a Storage Node from an Archive Node takes longer than restoring copies from other Storage Nodes because of the latency associated with retrievals from external archival storage systems.

About the `repair-data` script

To restore object data, you run the `repair-data` script. This script begins the process of restoring object data and works with ILM scanning to ensure that ILM rules are met.

Select **Replicated data** or **Erasure-coded (EC) data** below to learn the different options for the `repair-data` script, based on whether you are restoring replicated data or erasure-coded data. If you need to restore both types of data, you must run both sets of commands.

For more information about the `repair-data` script, enter `repair-data --help` from the command line of the primary Admin Node.

Replicated data

Two commands are available for restoring replicated data, based on whether you need to repair the entire node or only certain volumes on the node:

```
repair-data start-replicated-node-repair

repair-data start-replicated-volume-repair
```

You can track repairs of replicated data with this command:

```
repair-data show-replicated-repair-status
```

Erasure coded (EC) data

Two commands are available for restoring erasure-coded data, based on whether you need to repair the entire node or only certain volumes on the node:

```
repair-data start-ec-node-repair

repair-data start-ec-volume-repair
```

You can track repairs of erasure-coded data with this command:

```
repair-data show-ec-repair-status
```

Repairs of erasure-coded data can begin while some Storage Nodes are offline. However, if all erasure-coded data can't be accounted for, the repair can't be completed. Repair will complete after all nodes are available.

The EC repair job temporarily reserves a large amount of storage. Storage alerts might be triggered, but will resolve when the repair is complete. If there is not enough storage for the reservation, the EC repair job will fail. Storage reservations are released when the EC repair job completes, whether the job failed or succeeded.

Find hostname for Storage Node

1. Log in to the primary Admin Node:
   a. Enter the following command: `ssh admin@primary_Admin_Node_IP`
   b. Enter the password listed in the `Passwords.txt` file.
   c. Enter the following command to switch to root: `su -`
   d. Enter the password listed in the `Passwords.txt` file.

      When you are logged in as root, the prompt changes from `$` to `#`.

2. Use the `/etc/hosts` file to find the hostname of the Storage Node for the restored storage volumes. To see a list of all nodes in the grid, enter the following: `cat /etc/hosts`.

Repair data if all volumes have failed

If all storage volumes have failed, repair the entire node. Follow the instructions for **replicated data**, **erasure-coded (EC) data**, or both, based on whether you use replicated data, erasure-coded (EC) data, or both.

If only some volumes have failed, go to Repair data if only some volumes have failed.

> You can't run `repair-data` operations for more than one node at the same time. To recover multiple nodes, contact technical support.

Replicated data

If your grid includes replicated data, use the `repair-data start-replicated-node-repair` command with the `--nodes` option, where `--nodes` is the hostname (system name), to repair the entire Storage Node.

This command repairs the replicated data on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-node-repair --nodes SG-DC-SN3
```

> As object data is restored, the **Objects Lost** alert is triggered if the StorageGRID system can't locate replicated object data. Alerts might be triggered on Storage Nodes throughout the system. You should determine the cause of the loss and if recovery is possible. See Investigate lost objects.

Erasure coded (EC) data

If your grid contains erasure-coded data, use the `repair-data start-ec-node-repair` command with the `--nodes` option, where `--nodes` is the hostname (system name), to repair the entire Storage Node.

This command repairs the erasure-coded data on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-node-repair --nodes SG-DC-SN3
```

The operation returns a unique `repair ID` that identifies this `repair_data` operation. Use this `repair ID` to track the progress and result of the `repair_data` operation. No other feedback is returned as the recovery process completes.

> Repairs of erasure-coded data can begin while some Storage Nodes are offline. Repair will complete after all nodes are available.

Repair data if only some volumes have failed

If only some of the volumes have failed, repair the affected volumes. Follow the instructions for **replicated data**, **erasure-coded (EC) data**, or both, based on whether you use replicated data, erasure-coded (EC) data, or both.

If all volumes have failed, go to Repair data if all volumes have failed.

Enter the volume IDs in hexadecimal. For example, `0000` is the first volume and `000F` is the sixteenth volume. You can specify one volume, a range of volumes, or multiple volumes that aren't in a sequence.

All the volumes must be on the same Storage Node. If you need to restore volumes for more than one Storage Node, contact technical support.

Replicated data

If your grid contains replicated data, use the `start-replicated-volume-repair` command with the `--nodes` option to identify the node (where `--nodes` is the hostname of the node). Then add either the `--volumes` or `--volume-range` option, as shown in the following examples.

Single volume: This command restores replicated data to volume `0002` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volumes 0002
```

Range of volumes: This command restores replicated data to all volumes in the range `0003` to `0009` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volume-range
0003,0009
```

Multiple volumes not in a sequence: This command restores replicated data to volumes `0001`, `0005`, and `0008` on a Storage Node named SG-DC-SN3:

```
repair-data start-replicated-volume-repair --nodes SG-DC-SN3 --volumes
0001,0005,0008
```

> As object data is restored, the **Objects Lost** alert is triggered if the StorageGRID system can't locate replicated object data. Alerts might be triggered on Storage Nodes throughout the system. Note the alert description and recommended actions to determine the cause of the loss and if recovery is possible.

Erasure coded (EC) data

If your grid contains erasure-coded data, use the `start-ec-volume-repair` command with the `--nodes` option to identify the node (where `--nodes` is the hostname of the node). Then add either the `--volumes` or `--volume-range` option, as shown in the following examples.

Single volume: This command restores erasure-coded data to volume `0007` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volumes 0007
```

Range of volumes: This command restores erasure-coded data to all volumes in the range `0004` to `0006` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volume-range 0004,0006
```

Multiple volumes not in a sequence: This command restores erasure-coded data to volumes `000A`, `000C`, and `000E` on a Storage Node named SG-DC-SN3:

```
repair-data start-ec-volume-repair --nodes SG-DC-SN3 --volumes 000A,000C,000E
```

The `repair-data` operation returns a unique `repair ID` that identifies this `repair_data` operation. Use this `repair ID` to track the progress and result of the `repair_data` operation. No other feedback is returned as the recovery process completes.

Repairs of erasure-coded data can begin while some Storage Nodes are offline. Repair will complete after all nodes are available.

Monitor repairs

Monitor the status of the repair jobs, based on whether you use **replicated data**, **erasure-coded (EC) data**, or both.

You can also monitor the status of volume restoration jobs in process and view a history of restoration jobs completed in Grid Manager.

Replicated data

- To get an estimated percent completion for the replicated repair, add the `show-replicated-repair-status` option to the repair-data command.

  ```
  repair-data show-replicated-repair-status
  ```

- To determine if repairs are complete:
  1. Select **NODES** > ***Storage Node being repaired*** > **ILM**.
  2. Review the attributes in the Evaluation section. When repairs are complete, the **Awaiting - All** attribute indicates 0 objects.
- To monitor the repair in more detail:
  1. Select **SUPPORT** > **Tools** > **Grid topology**.
  2. Select ***grid*** > ***Storage Node being repaired*** > **LDR** > **Data Store**.
  3. Use a combination of the following attributes to determine, as well as possible, if replicated repairs are complete.

     Cassandra inconsistencies might be present, and failed repairs aren't tracked.

     - **Repairs Attempted (XRPA)**: Use this attribute to track the progress of replicated repairs. This attribute increases each time a Storage Node tries to repair a high-risk object. When this attribute does not increase for a period longer than the current scan period (provided by the **Scan Period — Estimated** attribute), it means that ILM scanning found no high-risk objects that need to be repaired on any nodes.

       High-risk objects are objects that are at risk of being completely lost. This does not include objects that don't satisfy their ILM configuration.

     - **Scan Period — Estimated (XSCM)**: Use this attribute to estimate when a policy change will be applied to previously ingested objects. If the **Repairs Attempted** attribute does not increase for a period longer than the current scan period, it is probable that replicated repairs are done. Note that the scan period can change. The **Scan Period — Estimated (XSCM)** attribute applies to the entire grid and is the maximum of all node scan periods. You can query the **Scan Period — Estimated** attribute history for the grid to determine an appropriate time frame.

Erasure coded (EC) data

To monitor the repair of erasure-coded data and retry any requests that might have failed:

1. Determine the status of erasure-coded data repairs:
   - Select **SUPPORT** > **Tools** > **Metrics** to view the estimated time to completion and the completion percentage for the current job. Then, select **EC Overview** in the Grafana section. Look at the **Grid EC Job Estimated Time to Completion** and **Grid EC Job Percentage Completed** dashboards.
   - Use this command to see the status of a specific `repair-data` operation:

     ```
     repair-data show-ec-repair-status --repair-id repair ID
     ```

   - Use this command to list all repairs:

```
repair-data show-ec-repair-status
```

The output lists information, including `repair ID`, for all previously and currently running repairs.

2. If the output shows that the repair operation failed, use the `--repair-id` option to retry the repair.

   This command retries a failed node repair, using the repair ID 6949309319275667690:

   ```
   repair-data start-ec-node-repair --repair-id 6949309319275667690
   ```

   This command retries a failed volume repair, using the repair ID 6949309319275667690:

   ```
   repair-data start-ec-volume-repair --repair-id 6949309319275667690
   ```

Check storage state after recovering Storage Node system drive

After recovering the system drive for a Storage Node, you must verify that the desired state of the Storage Node is set to online and ensure that the state will be online by default whenever the Storage Node server is restarted.

Before you begin

- You are signed in to the Grid Manager using a supported web browser.
- The Storage Node has been recovered, and data recovery is complete.

Steps

1. Select **SUPPORT** > **Tools** > **Grid topology**.
2. Check the values of **Recovered Storage Node** > **LDR** > **Storage** > **Storage State — Desired** and **Storage State — Current**.

   The value of both attributes should be Online.

3. If the Storage State — Desired is set to Read-only, complete the following steps:
   a. Click the **Configuration** tab.
   b. From the **Storage State — Desired** drop-down list, select **Online**.
   c. Click **Apply Changes**.
   d. Click the **Overview** tab and confirm that the values of **Storage State — Desired** and **Storage State — Current** are updated to Online.

Restore object data using Grid Manager

You can restore object data for a failed storage volume or Storage Node using Grid Manager. You can also use Grid Manager to monitor restoration processes in progress and display a restoration history.

Before you begin

- You have completed either of these procedures to format failed volumes:
  - Remount and reformat appliance storage volumes (manual steps)

    - Remount and reformat storage volumes (manual steps)
- You have confirmed that the Storage Node where you are restoring objects has a Connection State of **Connected** on the **NODES** > **Overview** tab in the Grid Manager.
- You have confirmed the following:
    - A grid expansion to add a Storage Node is not in process.
    - A Storage Node decommission is not in process or failed.
    - A recovery of a failed storage volume is not in process.
    - A recovery of a Storage Node with a failed system drive is not in process.
    - An EC rebalance job is not in process.
    - Appliance node cloning is not in process.

About this task

After you have replaced the drives and performed the manual steps to format the volumes, Grid Manager displays the volumes as candidates for restoration on the **MAINTENANCE** > **Volume restoration** > **Nodes to restore** tab.

Whenever possible, restore object data using the **Volume restoration** page in the Grid Manager. Follow these guidelines:

- If the volumes are listed at **MAINTENANCE** > **Volume restoration** > **Nodes to restore**, restore object data as described in the steps below. The volumes will be listed if:
    - Some, but not all, storage volumes in a node have failed
    - All storage volumes in a node have failed and are being replaced with the same number of volumes or more volumes

  The Volume restoration page in the Grid Manager also allows you to monitor the volume restoration process and view restoration history.
- If the volumes aren't listed in the Grid Manager as candidates for restoration, follow the appropriate steps for using the `repair-data` script to restore object data:
    - Restoring object data to storage volume (system drive failure)
    - Restore object data to storage volume where system drive is intact
    - Restore object data to storage volume for appliance

  If the recovered Storage Node contains fewer volumes than the node it is replacing, you must use the `repair-data` script.

You can restore two types of object data:

- Replicated data objects are restored from other locations, assuming that the grid's ILM rules were configured to make object copies available.
    - If an ILM rule was configured to store only one replicated copy and that copy existed on a storage volume that failed, you will not be able to recover the object.
    - If the only remaining copy of an object is in a Cloud Storage Pool, StorageGRID must issue multiple requests to the Cloud Storage Pool endpoint to restore object data.
    - If the only remaining copy of an object is on an Archive Node, object data is retrieved from the Archive

Node. Restoring object data to a Storage Node from an Archive Node takes longer than restoring object copies from other Storage Nodes.

- Erasure coded (EC) data objects are restored by reassembling the stored fragments. Corrupt or lost fragments are recreated by the erasure-coding algorithm from the remaining data and parity fragments.

> Volume restoration is dependent on the availability of resources where object copies are stored. Progress of volume restoration is nonlinear and might take days or weeks to complete.

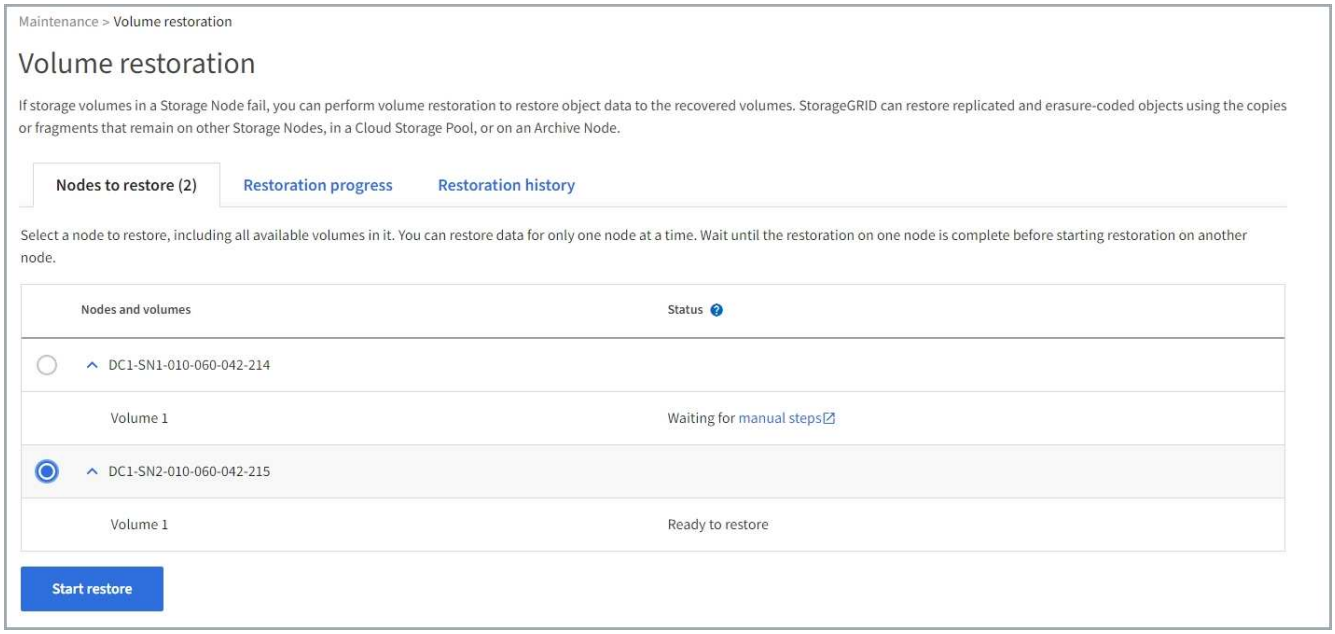
Restore failed volume or node

Follow these steps to restore a failed volume or node.

Steps

1. In Grid Manager go to **MAINTENANCE** > **Volume restoration**.
2. Select the **Nodes to restore** tab.

   The number on the tab indicates the number of nodes with volumes requiring restoration.



3. Expand each node to see the volumes in it that need restoration and their status.
4. Correct any issues preventing restoration of each volume that are indicated when you select Waiting for manual steps, if it displays as the volume status.
5. Select a node to restore where all the volumes indicate a Ready to restore status.

   You can only restore the volumes for one node at a time.

   Each volume in the node must indicate that it is ready to restore.

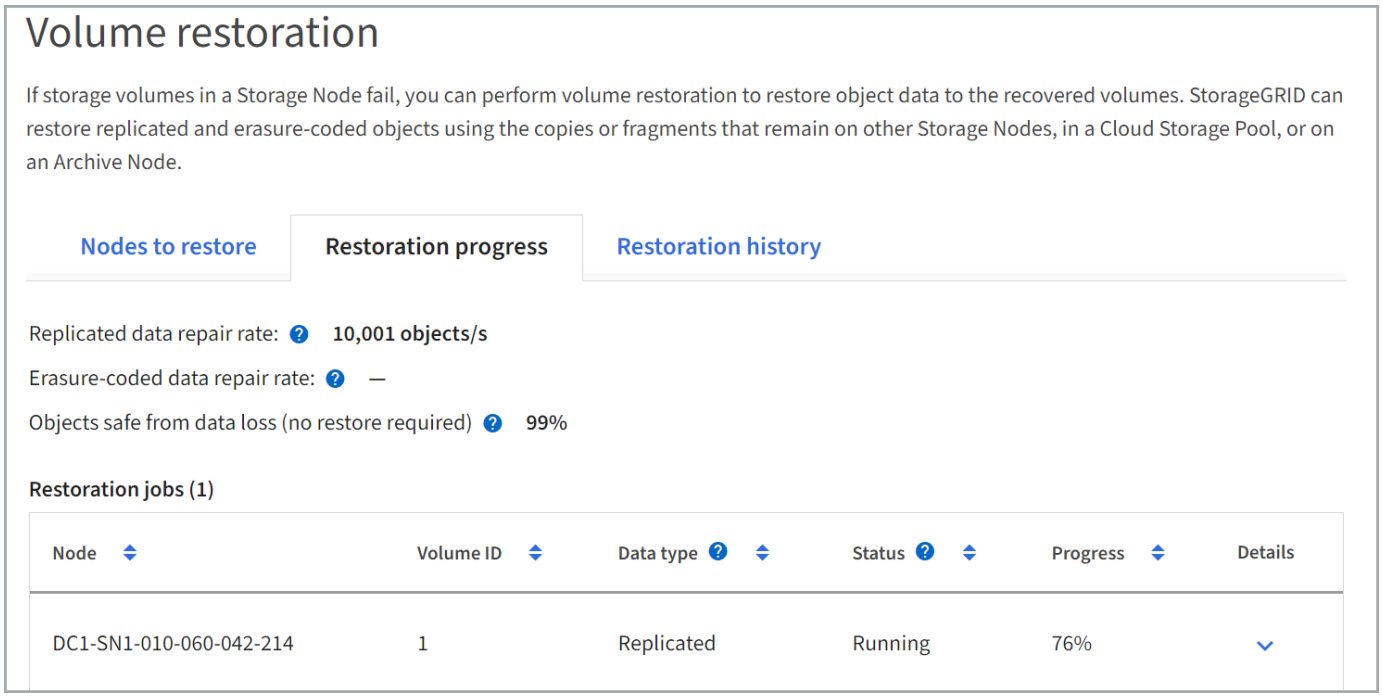
6. Select **Start restore**.
7. Address any warnings that might appear or select **Start anyway** to ignore the warnings and start the restoration.

Nodes are moved from the **Nodes to restore** tab to the **Restoration progress** tab when the restoration starts.

If a volume restoration can't be started, the node returns to the **Nodes to restore** tab.

View restoration progress

The **Restoration progress** tab shows the status of the volume restoration process and information about the volumes for a node being restored.



Data repair rates for replicated and erasure-coded objects in all volumes are averages summarizing all restorations in process, including those restorations initiated using the `repair-data` script. The percentage of objects in those volumes that are intact and don't require restoration is also indicated.

> Replicated data restoration is dependent on the availability of resources where the replicated copies are stored. Progress of replicated data restoration is nonlinear and might take days or weeks to complete.

The Restoration jobs section displays information about volume restorations started from Grid Manager.

- The number in the Restoration jobs section heading indicates the number of volumes that are either being restored or queued for restoration.
- The table displays information about each volume in a node being restored and its progress.
  - The progress for each node displays the percentage for each job.
  - Expand the Details column to display the restoration start time and job ID.
- If a volume restoration fails:
  - The Status column indicates failed.
  - An error appears, indicating the cause of failure.

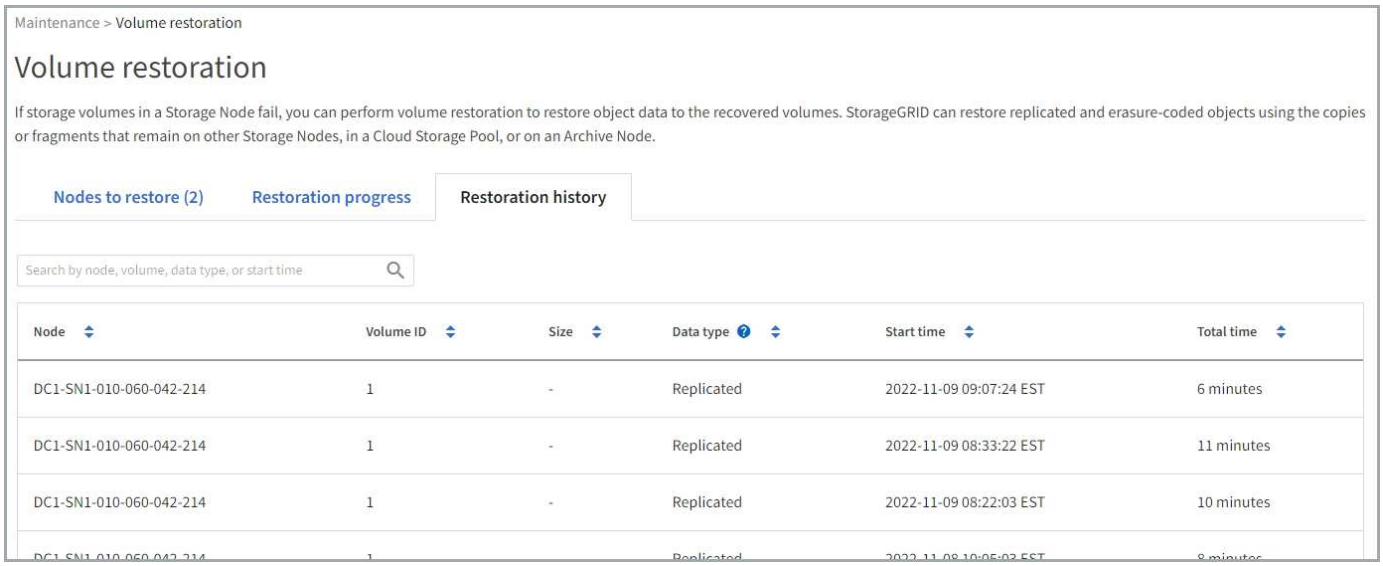
    Correct the issues indicated in the error. Then select **Retry** to re-initiate the volume restoration.

    If multiple restoration jobs have failed, selecting **Retry** starts the most recently failed job.

View restoration history

The **Restoration history** tab shows information about all volume restorations that have successfully completed.

Sizes aren't applicable for replicated objects and display only for restorations that contain erasure coded (EC) data objects.



Monitor repair-data jobs

You can monitor the status of repair jobs by using the `repair-data` script from the command line.

These include jobs that you initiated manually, or jobs that StorageGRID initiated automatically as part of a decommission procedure.

If you are running volume restoration jobs, monitor the progress and view a history of those jobs in the Grid Manager instead.

Monitor the status of `repair-data` jobs based on whether you use **replicated data**, **erasure-coded (EC) data**, or both.

Replicated data

- To get an estimated percent completion for the replicated repair, add the `show-replicated-repair-status` option to the repair-data command.

  ```
  repair-data show-replicated-repair-status
  ```

- To determine if repairs are complete:
  1. Select **NODES** > ***Storage Node being repaired*** > **ILM**.
  2. Review the attributes in the Evaluation section. When repairs are complete, the **Awaiting - All** attribute indicates 0 objects.
- To monitor the repair in more detail:
  1. Select **SUPPORT** > **Tools** > **Grid topology**.
  2. Select ***grid*** > ***Storage Node being repaired*** > **LDR** > **Data Store**.
  3. Use a combination of the following attributes to determine, as well as possible, if replicated repairs are complete.

     Cassandra inconsistencies might be present, and failed repairs aren't tracked.

     - **Repairs Attempted (XRPA)**: Use this attribute to track the progress of replicated repairs. This attribute increases each time a Storage Node tries to repair a high-risk object. When this attribute does not increase for a period longer than the current scan period (provided by the **Scan Period — Estimated** attribute), it means that ILM scanning found no high-risk objects that need to be repaired on any nodes.

       High-risk objects are objects that are at risk of being completely lost. This does not include objects that don't satisfy their ILM configuration.

     - **Scan Period — Estimated (XSCM)**: Use this attribute to estimate when a policy change will be applied to previously ingested objects. If the **Repairs Attempted** attribute does not increase for a period longer than the current scan period, it is probable that replicated repairs are done. Note that the scan period can change. The **Scan Period — Estimated (XSCM)** attribute applies to the entire grid and is the maximum of all node scan periods. You can query the **Scan Period — Estimated** attribute history for the grid to determine an appropriate time frame.

Erasure coded (EC) data

To monitor the repair of erasure-coded data and retry any requests that might have failed:

1. Determine the status of erasure-coded data repairs:
   - Select **SUPPORT** > **Tools** > **Metrics** to view the estimated time to completion and the completion percentage for the current job. Then, select **EC Overview** in the Grafana section. Look at the **Grid EC Job Estimated Time to Completion** and **Grid EC Job Percentage Completed** dashboards.
   - Use this command to see the status of a specific `repair-data` operation:

     ```
     repair-data show-ec-repair-status --repair-id repair ID
     ```

   - Use this command to list all repairs:

```
repair-data show-ec-repair-status
```

The output lists information, including `repair ID`, for all previously and currently running repairs.

2. If the output shows that the repair operation failed, use the `--repair-id` option to retry the repair.

This command retries a failed node repair, using the repair ID 6949309319275667690:

```
repair-data start-ec-node-repair --repair-id 6949309319275667690
```

This command retries a failed volume repair, using the repair ID 6949309319275667690:

```
repair-data start-ec-volume-repair --repair-id 6949309319275667690
```

Copyright information

Copyright © 2024 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.