IBM Cloud Object Storage System™
Version 3.14.7

Rebuilder Utility Guide

This edition applies to IBM Cloud Object Storage System and is valid until replaced by new editions.

Contents

Document information

Intended Purpose and Audience

This guide explains how a IBM Cloud Object Storage System™ maintains data integrity.

Chapter 1. Data challenges on a IBM Cloud Object Storage System™

Traditional storage systems need to protect data on three levels:

- Drives
- Nodes
- Geographies

This has been managed through:

- RAID arrays for drives
- High Availability systems or appliances for nodes
- Replication of data across geographies

A system protects against all simultaneously. This requires a unique solution to data protection. IBM achieves this using a group of software agents collectively known as the Rebuilder.

Chapter 2. Agents

The software agents that comprise the Rebuilder include:

`Scanning Agent`
Checks the consistency of the Slice names and revision numbers held by the Slicestor nodes

`Rebuild Agent`
Retrieves the new Slice data from other Slicestor nodes to repair a Slice on a Slicestor node

`Integrity Agent`
Checks the integrity of Slices on the Slicestor nodes

`Synchronization Agent`
Ensures that every Slice exists on both sides of a Vault Mirror

Note: This process is described in *Synchronizing Objects* in the *Vault Mirror Deployment Guide*.

Scanning agent

All Slicestor nodes scan for missing Slices and limit the number of listing requests processed at one time to throttle the scanning operations.

Due to the cooperative nature of scanning, Slicestor appliances do not list areas that other Slicestor appliances are scanning actively. A full scan of the IBM Cloud Object Storage System™ should take 48 hours. This means that any issue with a Slice is detected within 48 hours.

To maintain a high level of availability, the system should allow read and write operations to continue despite one or more Slicestor appliances being unreachable. Due to the distributed nature of the system and the massive amount of data involved, there is no centralized index that can inform a Rebuild agent as to which slices were written during an outage.

To ensure that drive failures and further outages do not render data unreadable, the Scanning Agent requests information about what slice names currently reside on each Slicestor appliance. To allow this process to continue even when certain Slicestor appliances are unable to scan, Scanning Agents coordinate in a distributed fashion. When Slicestor appliances receive requests about certain slice names, they record that these names have been scanned. They use this information to scan any Slice names that have yet to be scanned. This ensures that missing Slices are found quickly and Slicestor appliances do not duplicate work. When the Slicestor appliances find that together they have checked every Slice stored on the system, the process restarts.

The Scanning Agent uses this data to determine which any action should be taken for a given Object:

- If all Slices are present, nothing needs to happen.
- If a Slice is missing, reconstruct that Slice from the remaining Slices and upload the reconstructed Slice to the Slicestor appliance from which it is missing.
- If Slices cannot be reconstructed (if a client performed a delete operation during an outage); deprecate the existing Slices from the system to ensure that no storage leaks occur.

Rebuild agents

Individual Rebuild Agents address scenarios where Slices are missing from their respective Slicestor nodes or which are corrupted on a drive.

Each Slicestor appliance in a given Storage Pool ensures data integrity across all of the Slicestor appliances which reside in that Storage Pool. Even if a Slicestor appliance goes offline and misses some newly written Slice, the other Slicestor appliances notice that the Slice was missed and rebuild any Slices missing from the Slicestor appliance when it comes back online. These Slicestor nodes read this content of this Slice from Slicestor appliances without missing Slices and recreate the missing Slice on the recovered Slicestor appliance.

Rebuild Agents minimize impact on normal I/O operations to maintain expected system performance levels even during high amounts of rebuild activity. Rebuild activity is scheduled intelligently, so when I/O utilization drops, rebuilding becomes more aggressive, enhancing reliability without affecting performance.

Rebuild Agents operate continuously on all Slicestor appliances. To ensure a fair balance between rebuild and normal client I/O operations, adaptive algorithms within the Rebuilder continuously sample their impact to performance and reduce the rebuilding rate when they detect adverse impact to client performance.

Note: In general, the defined parameters together with the adaptive strategy provide desirable results and do not require tweaking. Discuss creating specific rebuilding optimization with your IBM Customer Success Engineer.

```
Example: Rebuilder Adaptive System Behavior

The Rebuilder's adaptive behavior changes in response to client I/O operations (indicated
in green). When client operations increase, the Rebuilder decreases. When client
operations reduces activity, the Rebuilder increases its activity as much as possible
without negatively impacting to client operations. The Rebuilder samples performance
many times a second to respond to changes in client operations.

In the figure below, when I/O increases suddenly, the rebuilding rate decreases rapidly,
to enable the client activity to achieve its maximum level of I/O.
```

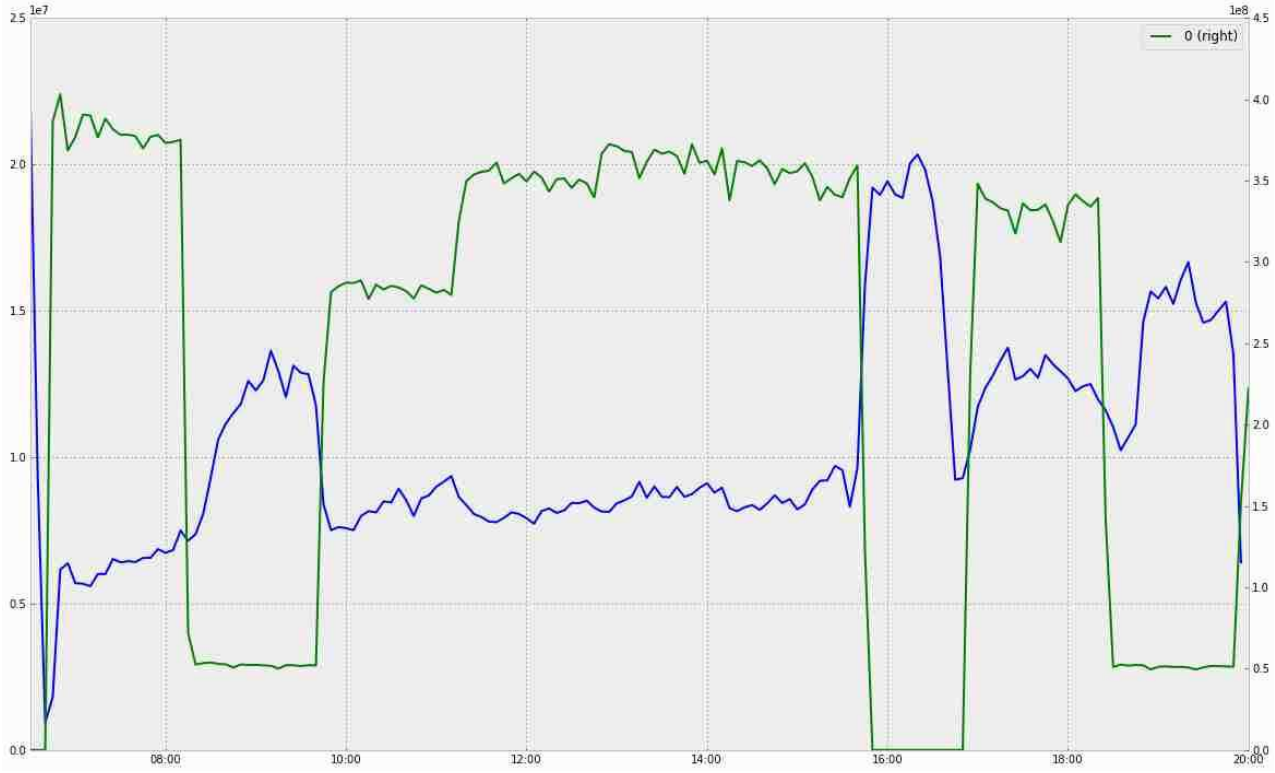

Figure 1. An illustration of how rebuilding rate (blue) adapts to changes of client rate (green).

Integrity agent

A major problem in many storage systems today is silent data corruption. A number of things can cause silent data corruption, including, but not limited to, drive controller errors or classes of errors occurring on the data bus.

The major problem with silent data corruption is that the host may never know if the data is corrupt unless it actively reads and checks for block-level data integrity, which most file systems do not.

To combat this problem, each Slicestor node runs an Integrity Agent which reads every Slice stored on that appliance and compares it against a 4-byte CRC32 checksum which is stored alongside that Slice. If the checksum does not match, the Slice is deprecated and then rebuilt from the other Slices for that Object per Slicestor node. This check occurs every time a client system attempts to read an Object. If the check fails, the Object and its component Slices are rebuilt.

Chapter 3. Bandwidth usage

Overview

Rebuilding is a necessary cost the IBM Cloud Object Storage System™ must incur to ensure reliability and availability over time.

While rebuilding has a minor impact on normal operations, it requires some bandwidth to keep the system healthy. The amount of bandwidth rebuilding consumes depends on a number of factors:

- Annual Data Rebuild Rate
- Total Annual Rebuild Network Traffic
- Individual Annual Slicestor Rebuild Network Traffic

Annual data rebuild rate

The most conservative (largest) amount of data to be rebuilt on annually is:

Annual data rebuild rate

$$D_B = (C_d \times C_{U\%} \times F_{Ry\%} \times d)$$

Table 1. Defined Variables for Annual Data Rebuild Rate Equation

Variable	Term	Definition	Example
C_d	Total Drive Capacity	Total capacity of each drive in the system.	If each drives has a capacity of 4 TB, then the Total Drive Capacity is `4 TB`.
$C_{U\%}$	Percent Storage Used	Average fill ratio of the system.	If the system is only 80% full, then the Percent Storage Used would be 80% or `0.8`.
$F_{Ry\%}$	Annual Unexpected Failure Rate	A conservative (high) estimation of the annual failure rate of drive drives is 4%. Due to the failing drive migration feature, drive failures do not necessarily result in a loss of data on that drive; often much of the data can be saved. By monitoring drive health statistics failing drives can be identified prior to failure over 75% of the time. Therefore, less than one in four failures is unexpected and requires rebuilding for recovery. This is because data is preemptively migrated off of drives suspected to fail soon. Therefore the annual unexpected drive failure rate is only a quarter of the annual drive failure rate, or 1% if the annual failure rate is 4%.	
`d`	Drive Count	This is the total number of drives in the system.	If the system is comprised of 100 Slicestor appliances with 45 drives each, then the Drive Count is `4,500`.

```
Example: Annual data rebuild rate

Assuming a drive capacity of 4 TB, fill ratio of 80%, annual unexpected failure rate
of 1% and 2,500 drives (a system with 10 PB of raw drive storage), the annual data to
rebuild is:

80 TB = (4 TB × 80% × 1% × 2500)
```

Total annual rebuild network traffic

With this information the expected rebuild traffic can be calculated. For every byte of data lost, threshold bytes will have to be read to rebuild it. Thus the rebuild traffic on a system per year is:

Annual Rebuild Network Traffic

$B_{Dy} = (D_B \times T)$

Table 2. Defined Variables for Annual Rebuild Network Traffic Equation

Variable	Term	DefinitionOn the contrary
D_B	Annual Data to Rebuild	Result of Annual Data Rebuild Rate.
T	IDA Threshold	How many Slices are required to recover a piece of data

```
Example: Total annual Slicestor rebuild network traffic

So 80 TB of data must be rebuilt each year. This works out to 2.65 MBps. The total amount
of traffic required to rebuild 80 TB of lost data is this amount multiplied by the threshold.
If the threshold is assumed to be 20:

Example of Annual Amount of Rebuild Network Traffic

1600 TB = (80 TB × 20)
```

Individual annual Slicestor rebuild network traffic

Since cooperative rebuilding decentralizes the process of rebuilding, this load is distributed evenly across the Slicestor appliances.

To find the amount of network traffic a single Slicestor appliance uses:

Individual Annual Rebuild Network Traffic

$B_{Dny} = (B_{Dy} \div n_s)$

```
Example: Computing the individual annual Slicestor rebuild network traffic

Rebuilding requires 1.6 PB of network traffic annually or 53.16 MBps. If each Slicestor
appliance holds 20 drives, there would be 125 Slicestor appliances. Knowing this, the annual
amount of rebuild traffic per Slicestor appliance can be computed.

2.8 TB = (1600 TB ÷ 125)

Each 125 Slicestor appliance uses 12.8 TB of traffic annually for rebuilding operations
or a per-appliance average of 435.52 KBps.
```

Tip: To convert annual to per second, divide by the annual amount by 31,557,600.

Chapter 4. Prioritized rebuilder

Theory of operation

The Prioritized Rebuilder uses a distributed index as a priority queue. An index is used as a priority queue. Entries can be removed by name from the priority queue. The first N highest priority items are leased from the queue.

There is one priority queue per Object Vault. The queue is stored as an index on the object vault. Also, the queue is divided over each set to eliminate cross-set traffic.

To determine the priority of rebuilding a given source on a given vault, use the number of missing slices and the IDA of the vault itself.

Scanning agents feed work into the shared priority queue for each vault.

Rebuild agents pull work from the shared prioritized queue based on rebuild need for each vault. A list is kept of all relevant vaults and their priority.

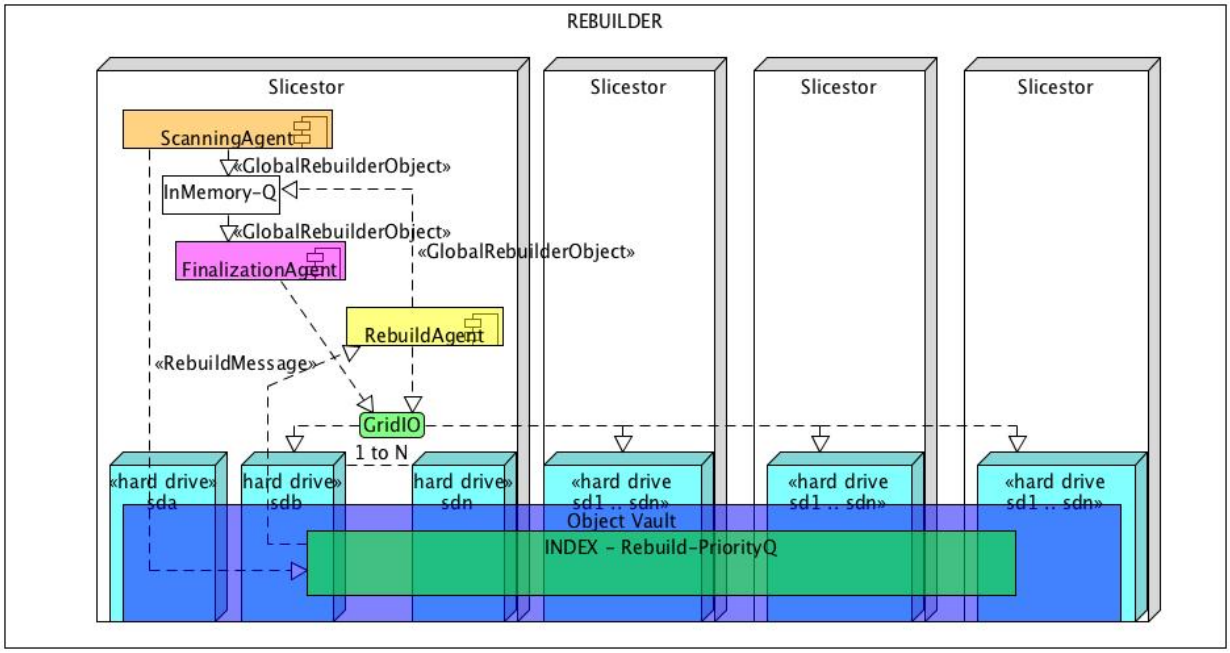

Figure 2. Rebuilder

Scanning agent

Scanning agents use the Rebuilder's cooperative scanning mechanism to share the scanning of the name range for rebuildable sources. Instead of feeding a local rebuild-agent mailbox with results of the scans, the scanning agents use the information to populate the rebuild queue. A scanning agents decides the priority of the Rebuilder work object in the rebuild queue based on the number of slices that can be lost before data loss. In event of a failure to access the rebuild queue, the Scanning agent populates a local in memory FIFO queue with rebuild work objects.

If a source is missing slices less than or equal to a missing slice threshold of missing slices, the rebuild object will be added to the **localrebuild** queue. If a source is missing more slices than the missing slice threshold, it will be added to the index as a single entry.

Rebuild agent

Rebuild agents take work from the Rebuild Work Queue (Index), pull work from the Local Rebuilder Queues, and work on the most endangered source names first. Rebuild agents can be throttled by modifying the amount of Rebuilder work objects each one requests from the Rebuild queue. Rebuild agents can be throttled by the amount of rebuilding tasks they are able to run by adjusting the number of available permits available to the Rebuilder controller.

If a Rebuild agent begins work on a source name that cannot be rebuilt due to a Slicestor® device being down or a disk being quarantined, the rebuild object will be removed and retried when the object is scanned again.

In event of a failure to access the Rebuilder queue, the Rebuilder will process Rebuilder work objects populated in a local in-memory queue from the scanning agent. Rebuilder work objects from the index will be processed before items populated in the in-memory FIFO queue.

Rebuild queue

A Rebuild queue is represented by a shared prioritized index per vault per storage set containing a string key made up of the priority and the source name of the object to be rebuilt to reduce duplicates. The value in the prioritized index contains an object that represents the rebuild work needed. The rebuild object contains the missing slices and the version number relating to the source name of the object we are rebuilding as well as more information relating to rebuild.

A Rebuild queue is in the object vault, if available. A Rebuild queue creates a new storage type to enforce locality of index nodes within a set. This requires a vault format upgrade for the object vault.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM® in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com® are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at Copyright and trademark information at www.ibm.com/legal/copytrade.shtml.

Accesser®, Cleversafe®, ClevOS™, Dispersed Storage®, dsNet®, IBM Cloud Object Storage Accesser®, IBM Cloud Object Storage Dedicated™, IBM Cloud Object Storage Insight™, IBM Cloud Object Storage Manager™, IBM Cloud Object Storage Slicestor®, IBM Cloud Object Storage Standard™, IBM Cloud Object Storage System™, IBM Cloud Object Storage Vault™, SecureSlice™, and Slicestor® are trademarks or registered trademarks of Cleversafe, an IBM Company and/or International Business Machines Corp.

Other product and service names might be trademarks of IBM or other companies.

Homologation statement

This product may not be certified in your country for connection by any means whatsoever to interfaces of public telecommunications networks. Further certification may be required by law prior to making any such connection. Contact an IBM representative or reseller for any questions.

Printed in USA