CDP Public Cloud Management Console

Data Protection

Date published: 2019-08-22
Date modified:

CLOUDERA

Legal Notice

Contents

Ensuring data recovery and preventing data loss in AWS

Different recovery and preventive methods can be used to ensure recovery in case of accidental deletes and corruption of cloud storage data. Monitoring activities and the state of data can help determine what caused the data loss, and with the available backup and restore tools you can ensure the recovery of data stored in S3.

Monitoring activities and data

You can use CDP auditing, CloudTrail and S3 Storage Lens to monitor activities and events in the AWS S3 bucket.

CDP auditing

Server and audit logs can give a more detailed view of requests made to the AWS S3 bucket through CDP. The audit archiving can collect or log evidence of activities in a system. The audit and server logs are enabled by default from Cloudera Runtime 7.2.15. For more information, see the *Enable AWS S3 Access logs for server side logging* and the *Auditing documentation* in CDP Public Cloud.

CloudTrail

CloudTrail is integrated with Amazon S3, providing a record of actions performed by a user, role or AWS service. The events are captured as a subset of API calls. CloudTrail can be enabled to log the events happening on the S3 data. For more information, see the *Logging Amazon S3 API calls using AWS CloudTrail* documentation.

S3 Storage Lens

S3 Storage Lens provides an overview of the data stored in S3 and activity trends. S3 Storage Lens can be used to have a brief summary of what the state of the data is in S3 for various buckets. For more information, see the *S3 Storage Lens* introduction.

Related Information

Enable AWS S3 Access logs for server side logging

Auditing documentation

Logging Amazon S3 API calls using AWS CloudTrail

S3 Storage Lens

Backing up versions and objects

You can use S3 versioning and AWS Backup for S3 to create backup of data stored in S3.

S3 versioning

S3 versioning can keep multiple versions of the objects stored in the S3 bucket. Using the versioning, it is easier to recover the data that was deleted by accident or due to application failure. For more information, see the *Using versioning in S3 buckets* documentation.

As versioning is not enabled by default, Cloudera recommends enabling the S3 versioning to ensure the recovery of lost data using the *Enabling versioning on buckets* and *Working with versioned S3 buckets* documentation.

You can also find how to retrieve object versions from a bucket when versioning is enabled in the *Amazon S3* documentation. LifeCycle Rules can be used with versioning to manage the lifecycle of the objects in an S3 bucket. For more information, see the *Managing your storage lifecycle* documentation.

Ease of recovery

The version ID of the object is required when recovering an object to its previous version. In case the whole bucket or more than one object needs to be recovered, there must be a list of all the versions for each object and the recovery is based on the timestamp of the objects.

Recovery scenarios

If there is a request from a job to remove a certain file/folder, instead of actually removing the object, the current file must have a 'delete marker'. This 'delete marker' serves as a new version on top of the current version of the object, and the 'delete marker' is be deleted. In case of a job failure or accidental deletion, it is hard to decide if an object is deleted by accident or it was a 'delete marker' file. There is no way to reset all of the objects to a particular time before a job was run. In case someone deletes an older version permanently or a retention policy removes an older version, there is no way to recover the version.

Security and encryption

The same IAM roles are applied for the versioning as for the S3 bucket. Versions are available based on the access of the user to the bucket data. Versioned data also follows the same encryption mechanism as the S3 stored data, such as SSE, SSE-KMS, SSE-c, and so on.

Object level recovery

Versioning happens on an object level: versioning a bucket applies the versioning to all the objects, and recovery with versioning is per object level.

Backup scheduling and location

Lifecycle policies can be configured to manage the scheduling of backups. Only one version is considered as the current version, a new current version can be created by modifying the object. The versions are stored in the same bucket as the objects. These versions can be moved across different buckets. Versioning can be enabled for all regions.

AWS Backup for S3

AWS Backup offers backup management, policy-based and tag-based backup, lifecycle management policies, cross-region and cross-account backup features, and can be used with S3 versioning. When using AWS Backup, Cloudera recommends setting a lifecycle expiration period for the S3 versions as all unexpired versions of the S3 data will be backed up and stored, which can increase cloud cost. For more information about the AWS Backup features and availability, see the *What is AWS Backup* documentation.

Ease of recovery

Using AWS Backup, you can choose from a list of recovery points that indicate the state of S3 data at that point in time. The whole bucket can be restored or up to 5 prefixes to the recovery point's data. Data can be restored in the source bucket, another existing bucket or in a newly created bucket, but the bucket must be in the same region as the backup vault.

Recovery scenarios

If a certain file or folder is removed due to a job request, the bucket can be recovered to a point of time before the job started or if the prefix (or path folder) is identified, the file or folder can be restored from that previous recovery point.

If someone permanently deletes some versions and objects from the bucket, the previous recovery point can be used to recover the data. Separate IAM roles for backup and S3 can ensure separation in access to both data.

If someone deletes the recovery points, there is no way to recover that recovery point as the backup is incremental. Vault lock can be used to prevent the deletion of recovery points.

Security and encryption

Different IAM roles defined for backup vault and the S3 bucket can ensure different access permissions. A KMS key is required to encrypt all backed up data, which can be a KMS key you created or a default AWS one.

Object level recovery

AWS Backup allows up to 5 prefixes to recover objects or the whole bucket data is recovered.

AWS Backup creates a backup of all your S3 versions, but restores only the latest version from the version stack at any point in time. This AWS Backup limitation can be harmful when recovering from corrupted data. In this case, the corrupted version needs to be the 'delete marker' and restoring from the previous version, or permanently deleting the corrupted version and using the previous one as the current version.

Backup scheduling and location

Creating a backup plan allows you to schedule and prepare for backup when needed in the defined interval. However, on-demand backup can also be created beside the scheduled backup. A bucket can only be backed up in the supported regions of AWS backup. Restoring can be done to the same region bucket as well. AWS Backup is supported for Amazon S3 in all regions except China (Beijing), China (Ningxia), Europe (Spain), Europe (Zurich), Asia Pacific (Hyderabad), Asia Pacific (Melbourne) Region.

Comparison of S3 versioning and AWS Backup

Both S3 versioning and AWS Backup are easy to implement and use for backing up the S3 bucket and with the help of IAM roles a more sufficient security level can be configured for both tools. However, when using S3 versioning, there is no straight way to provide the whole bucket's previous state at a particular time, and an external script is required to restore prefixes/directories to a specific time. For AWS Backup, if a current version (non-delete marker) is present and a previous version needs to be restored, the restore job prefers the current version, so it could mean that in cases of data corruption we fall back to versioning as our solution to restore a particular version. You also need to enable the S3 versioning to use AWS Backup.

The S3 versioning and AWS Backup have different pricing based on usage. For more information, see the *AWS S3 Pricing* page.

Related Information
Using versioning in S3 buckets
Enabling versioning on buckets
Working with versioned S3 buckets
Amazon S3
Managing your storage lifecycle
What is AWS Backup
AWS S3 Pricing

Restoring versions and objects

You can use Point-in-time Restore and S3 pit restore when multiple files and larger directory trees need to be restored from versioned buckets.

Point-in-time Restore

The S3 Versioning and the Amazon EventBridge can restore individual objects or large datasets. This solution provides a complete bucket inventory and determines object changes from the specified timestamp, the snapshot can be restored into the same or a different bucket. For more information about the process and enablement of point-in-time restore, see *Point-in-time restore for Amazon S3 buckets*.

S3 pit restore Tool

The S3 pit restore Tool can restore previous versions of the data stored in S3 when S3 versioning is enabled. You can restore some or all of the files to a certain point in time to a local file system, the same S3 bucket, or a different S3 bucket. For more information, see the *S3 pit restore* documentation.

Comparison of Point-in-time Restore and S3 pit restore

Both Point-in-time Restore and S3 pit restore can be used to restore directories/prefixes to a previous version when S3 versioning is enabled. The Point-in-time Restore can also provide audit/event logs to show changes in S3, which give a more detailed view of the performed operations. However, the Point-in-time Restore can only restore changed objects and relies on other AWS services and requires more steps. The dependency on other AWS services also affects the pricing calculation for Point-in-time Restore. For more information, see *Point-in-time restore for Amazon S3 buckets* and the *AWS S3 Pricing* page. While the S3 pit restore Tool has no additional cost, the maintainability of the tool is not certain and requires users to configure their own access, secret key, and region. Furthermore, Delegation Tokens will not work.

Related Information

Point-in-time restore for Amazon S3 buckets
S3 pit restore
AWS S3 Pricing