

Cloudera Runtime 7.0.2

Cloud Data Access

Date published: 2019-09-23

Date modified:

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Cloud storage connectors overview.....	5
The Cloud Storage Connectors.....	5
Working with Amazon S3.....	6
Limitations of Amazon S3.....	7
Configuring Access to S3.....	7
Using EC2 Instance Metadata to Authenticate.....	8
Referencing S3 Data in Applications.....	8
Configuring Per-Bucket Settings.....	8
Customizing Per-Bucket Secrets Held in Credential Files.....	9
Configuring Per-Bucket Settings to Access Data Around the World.....	10
Encrypting Data on S3.....	10
SSE-S3: Amazon S3-Managed Encryption Keys.....	11
SSE-KMS: Amazon S3-KMS Managed Encryption Keys.....	11
SSE-C: Server-Side Encryption with Customer-Provided Encryption Keys.....	12
Configuring Encryption for Specific Buckets.....	13
Encrypting an S3 Bucket with Amazon S3 Default Encryption.....	13
Performance Impact of Encryption.....	14
Using S3Guard for Consistent S3 Metadata.....	14
Introduction to S3Guard.....	14
Configuring S3Guard.....	15
Monitoring and Maintaining S3Guard.....	16
Disabling S3Guard and destroying a table.....	16
Pruning Old Data from S3Guard Tables.....	16
Importing a Bucket into S3Guard.....	17
Verifying that S3Guard is Enabled on a Bucket.....	17
Using the S3Guard CLI.....	18
S3Guard: Operational Issues.....	18
Safely Writing to S3 Through the S3A Committers.....	19
Introducing the S3A Committers.....	19
Configuring Directories for Intermediate Data.....	20
Using the Directory Committer in MapReduce.....	21
Verifying That an S3A Committer Was Used.....	21
Cleaning up after failed jobs.....	21
Using the S3Guard Command to List and Delete Uploads.....	21
Advanced Committer Configuration.....	22
Securing the S3A Committers.....	23
The S3A Committers and Third-Party Object Stores.....	23
Limitations of the S3A Committers.....	23
Troubleshooting the S3A Committers.....	24
Security Model and Operations on S3.....	24
S3A and Checksums (Advanced Feature).....	25
A List of S3A Configuration Properties.....	26
Working with versioned S3 buckets.....	26
Working with Third-party S3-compatible Object Stores.....	27
Improving Performance for S3A.....	28

Working with S3 buckets in the same AWS region.....	28
Configuring and tuning S3A block upload.....	28
Optimizing S3A read performance for different file types.....	31
S3 Performance Checklist.....	32
Troubleshooting S3 and S3Guard.....	32

Cloud storage connectors overview

This content provides information and steps required for, using, securing, tuning performance, and troubleshooting access to the cloud storage services using CDP cloud storage connectors available for Amazon Web Services (Amazon S3). This content provides information and steps required for, using, securing, tuning performance, and troubleshooting access to the cloud storage services using CDP cloud storage connectors available for Amazon Web Services (Amazon S3) and Microsoft Azure (ADLS Gen 2).

The primary audience of this content are the administrators and users of CDP deployed on cloud Infrastructure-as-a-Service (IaaS) such as Amazon Web Services (AWS). The primary audience of this content are the administrators and users of CDP deployed on cloud Infrastructure-as-a-Service (IaaS) such as Amazon Web Services (AWS) and Microsoft Azure.

You may also use this content if your CDP is deployed in your own data center and you plan to access cloud storage through the connectors; however, your experience and performance may vary based on the network bandwidth between your data center and the cloud storage service. In addition, you can optionally configure other features where available.

The Cloud Storage Connectors

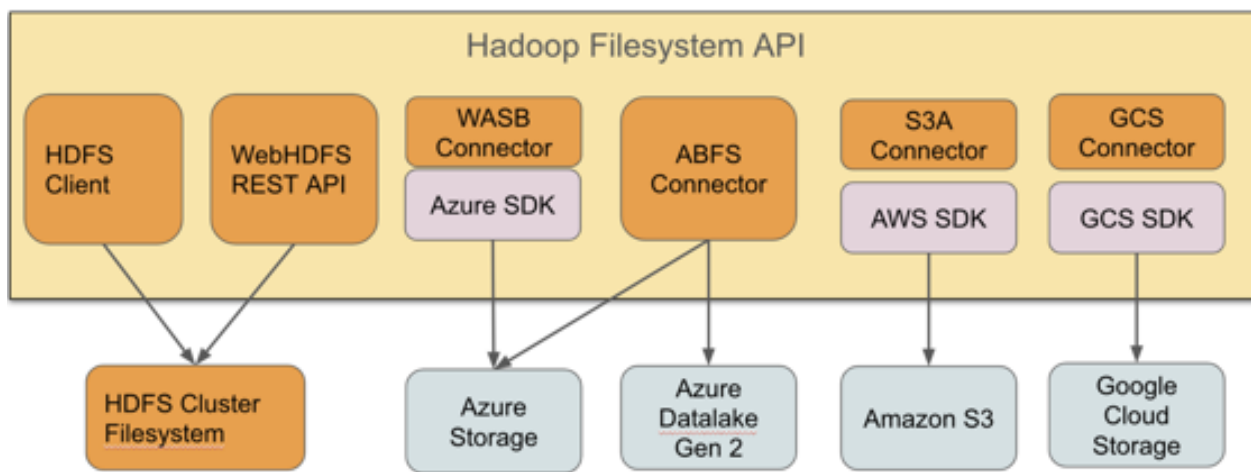
When deploying CDP clusters on cloud IaaS, you can take advantage of the native integration with the object storage services available on Amazon S3 on AWS. This integration is through cloud storage connectors included with CDP. Their primary function is to help you connect to, access, and work with data the cloud storage services.

The cloud connectors allow you to access and work with data stored in Amazon S3, including but not limited to the following use cases:

- Collect data for analysis and then load it into Hadoop ecosystem applications such as Hive or Spark directly from cloud storage services.
- Persist data to cloud storage services for use outside of CDP clusters.
- Copy data stored in cloud storage services to HDFS for analysis and then copy back to the cloud when done.
- Share data between multiple CDP clusters – and between various external non-CDP systems.
- Back up CDP clusters using distcp.

The cloud object store connectors are implemented as individual Hadoop modules. The libraries and their dependencies are automatically placed on the classpath.

Hadoop Ecosystem Applications (Hive queries, MapReduce, Spark, etc.)



Amazon S3 is an object store. The S3A connector implements the Hadoop filesystem interface using AWS Java SDK to access the web service, and provides Hadoop applications with a filesystem view of the buckets. Applications can manipulate data stored in Amazon S3 buckets with an URL starting with the `s3a://` prefix.

Amazon S3 can not be used as a replacement for HDFS as the cluster filesystem in CDP. Amazon S3 can be used as a source and destination of work.

Working with Amazon S3

The [Amazon S3](#) object store is the standard mechanism to store, retrieve, and share large quantities of data in AWS.

The features of Amazon S3 include:

- Object store model for storing, listing, and retrieving data.
- Support for objects up to 5 terabytes, with many petabytes of data allowed in a single "bucket".
- Data is stored in Amazon S3 in [buckets](#) which are stored in different [AWS regions](#).
- Buckets can be restricted to different users or [IAM roles](#).
- Data stored in an Amazon S3 bucket is billed based on the size of data how long it is stored, and on operations accessing this data. In addition, you are billed when you transfer data between regions:
 - Data transfers between an Amazon S3 bucket and a cluster running in the same region are free of download charges (except in the special case of buckets in which data is served on a user-pays basis).
 - Data downloaded from an Amazon S3 bucket located outside the region in which the bucket is hosted is billed per megabyte.
 - Data downloaded from an Amazon S3 bucket to any host over the internet is also billed per-Megabyte.
- Data stored in Amazon S3 can be backed up with [Amazon Glacier](#).

The Hadoop client to S3, called "S3A", makes the contents of a bucket appear like a filesystem, with directories, files in the directories, and operations on directories and files. As a result, applications which can work with data stored in HDFS can also work with data stored in S3. However, since S3 is an object store, it has certain limitations that you should be aware of.

Limitations of Amazon S3

Even though Hadoop's S3A client can make an S3 bucket appear to be a Hadoop-compatible filesystem, it is still an object store, and has some limitations when acting as a Hadoop-compatible filesystem.

The key things to be aware of are:

- Operations on directories are potentially slow and non-atomic.
- Not all file operations are supported. In particular, some file operations needed by Apache HBase are not available — so HBase cannot be run on top of Amazon S3.
- Data is not visible in the object store until the entire output stream has been written.
- Neither the per-file and per-directory permissions supported by HDFS nor its more sophisticated ACL mechanism are supported.
- Bandwidth between your workload clusters and Amazon S3 is limited and can vary significantly depending on network and VM load.

For these reasons, while Amazon S3 can be used as the source and store for persistent data, it cannot be used as a direct replacement for a cluster-wide filesystem such as HDFS, or be used as defaultFS.

Configuring Access to S3

IDBroker is a REST API built as part of Apache Knox's authentication services. It allows an authenticated user to exchange a set of credentials or a token for cloud vendor access tokens. It manages mapping LDAP users to FreeIPA cloud identities for data access. It performs identity mapping for access to object stores.

This section provides information about the options that are enabled by default. For information on how IDBroker works in CDP, see *Management Console* documentation.

For Apache Hadoop applications to be able to interact with Amazon S3, they must know the AWS access key and the secret key. This can be achieved in multiple ways, including configuration properties, environment variables, and EC2 instance metadata. While the first two options can be used when accessing S3 from a cluster running in your own data center. IAM roles, which use instance metadata should be used to control access to AWS resources if your cluster is running on EC2.

Deployment Scenario	Authentication Options
Cluster runs on EC2	Use IAM roles to bind your EC2 VMs to a specific role, and so manage its access to AWS resources. IDBroker will provide authenticated users with the AWS credentials they need to access data in S3. There is no need to manually provide any AWS login credentials directly to users
Cluster runs in your own data center	Use configuration properties to authenticate. You can set the configuration properties globally or per-bucket.

Temporary security credentials, also known as "session credentials", can be issued. These consist of a secret key with a limited lifespan, along with a session token, another secret which must be known and used alongside the access key. The secret key is never passed to AWS services directly. Instead it is used to [sign the URL and headers of the HTTP request](#).

By default, the S3A filesystem client follows the following authentication chain:

1. The AWS login details are looked for in the job configuration settings.
2. The AWS environment variables are then looked for.
3. An attempt is made to query the Amazon EC2 Instance Metadata Service to retrieve credentials published to EC2 VMs.

Using EC2 Instance Metadata to Authenticate

If your cluster is running on EC2, the standard way to manage access is via [Amazon Identity and Access Management \(IAM\)](#), which allows you to create users, groups, and roles to control access to services such as Amazon S3 via attached policies.

A role does not have any credentials such as password or access keys associated with it. Instead, if a user is assigned to a role, access keys are generated dynamically and provided to the user when needed. For more information, refer to [IAM Roles for Amazon EC2](#) in Amazon documentation.

When launching your cluster on EC2, specify an IAM role that you want to use; if you are planning to use S3 with your cluster, make sure that the role associated with the cluster includes a policy that grants access to S3. For more information, refer to [Using an IAM Role to Grant Permissions to Applications Running on Amazon EC2 Instances](#) in Amazon documentation. No additional configuration is required.



Note: You can use IAM Roles to control access to keys stored in Amazon's KMS Key Management service. For more information, refer to [Overview of Managing Access to Your AWS KMS Resources](#) in Amazon documentation.

Referencing S3 Data in Applications

You can reference data in Amazon S3 using a URL starting with the `s3a://` prefix followed by bucket name and path to file or directory.

The URL structure is:

```
s3a://<bucket>/<dir>/<file>
```

For example, to access a file called "mytestfile" in a directory called "mytestdir", which is stored in a bucket called "mytestbucket", the URL is:

```
s3a://mytestbucket/mytestdir/mytestfile
```

The following FileSystem shell commands demonstrate access to a bucket named mytestbucket:

```
hadoop fs -ls s3a://mytestbucket/  
hadoop fs -mkdir s3a://mytestbucket/testDir  
hadoop fs -put testFile s3a://mytestbucket/testFile  
hadoop fs -cat s3a://mytestbucket/testFile
```

Configuring Per-Bucket Settings

You can specify bucket-specific configuration values which override the common configuration values.

1. Authentication mechanisms and credentials
2. Encryption settings
3. The specific S3 endpoints to send HTTP requests to. This is essential when working with S3 regions which only support the "V4 authentication API", in case of which callers must always declare the explicit region.

All `fs.s3a` options other than a small set of unmodifiable values (currently `fs.s3a.impl`) can be set on a per-bucket basis.

To set a bucket-specific option:

1. Add a new configuration, replacing the `fs.s3a.` prefix on an option with `fs.s3a.bucket.BUCKETNAME`, where `BUCKETNAME` is the name of the bucket.

For example, if you are configuring access key for a bucket called "nightly", instead of using `fs.s3a.access.key` property name, use `fs.s3a.bucket.nightly.access.key`.

2. When connecting to a bucket, all options explicitly set for that bucket will override the base `fs.s3a.` values, but they will not be picked up by other buckets.

Example

You may have a base configuration to use the IAM role information available when deployed in Amazon EC2 VM or an AWS container service.

```
<property>
  <name>fs.s3a.aws.credentials.provider</name>
  <value>org.apache.hadoop.fs.s3a.auth.IAMInstanceCredentialsProvider</value>
</property>
```

This will be the default authentication mechanism for S3A buckets.

A bucket `s3a://nightly/` used for nightly data uses a session key, so its bucket-specific configuration is:

```
<property>
  <name>fs.s3a.bucket.nightly.access.key</name>
  <value>AKAACCE-SKEY-2</value>
</property>
<property>
  <name>fs.s3a.bucket.nightly.secret.key</name>
  <value>SESSION-SECRET-KEY</value>
</property>
<property>
  <name>fs.s3a.bucket.nightly.session.token</name>
  <value>SHORT-LIVED-SESSION-TOKEN</value>
</property>

<property>
  <name>fs.s3a.bucket.nightly.aws.credentials.provider</name>
  <value>org.apache.hadoop.fs.s3a.TemporaryAWSCredentialsProvider</value>
</property>
```

Finally, the public `s3a://landsat-pds/` bucket could be accessed anonymously, so its bucket-specific configuration is:

```
<property>
  <name>fs.s3a.bucket.landsat-pds.aws.credentials.provider</name>
  <value>org.apache.hadoop.fs.s3a.AnonymousAWSCredentialsProvider</value>
</property>
```

For all other buckets, the base configuration is used.

Customizing Per-Bucket Secrets Held in Credential Files

JCEKS credential files support the same per-bucket settings as those in XML files. To provide different credentials for different buckets, simply create per-bucket entries in the JCEKS file with the relevant secrets.

Example

1. Set base properties for `fs.s3a.secret.key` and `fs.s3a.access.key` in the JCEKS file. These will be the default.
2. Set all non-security properties per-bucket for a bucket called "frankfurt-1" in the `core-site.xml`. These will override the base properties when talking to the bucket "frankfurt-1".

- For the AWS authentication secrets, set the `fs.s3a.frankfurt-1.access.key` and `fs.s3a.frankfurt-1.secret.key` properties in the JCEKS file.

Related Information

[Credential Provider API Guide](#)

Configuring Per-Bucket Settings to Access Data Around the World

S3 buckets are hosted in different AWS regions, the default being "US-East". The S3A client talks to this region by default, issuing HTTP requests to the server `s3.amazonaws.com`. This central endpoint can be used for accessing any bucket in any region which supports using the V2 Authentication API, albeit possibly at a reduced performance.

Each region has its own S3 endpoint, [documented by Amazon](#). The S3A client supports these endpoints. While it is generally simpler to use the default endpoint, direct connections to specific regions (i.e. connections via region's own endpoint) may deliver performance and availability improvements, and are mandatory when working with the most recently deployed regions, such as Frankfurt and Seoul.

When deciding which endpoint to use, consider the following:

- Applications running in EC2 infrastructure do not pay for data transfers to or from local S3 buckets. In contrast, they will be billed for access to remote buckets. Therefore, wherever possible, always use local buckets and local copies of data.
- When the V1 request signing protocol is used, the default S3 endpoint can support data transfer with any bucket.
- When the V4 request signing protocol is used, AWS requires the explicit region endpoint to be used — hence S3A must be configured to use the specific endpoint. This is done in the configuration option `fs.s3a.endpoint`.
- All endpoints other than the default endpoint only support interaction with buckets local to that S3 instance.

If the wrong endpoint is used, the request may fail. This may be reported as a 301 redirect error, or as a 400 Bad Request. Take these failures as cues to check the endpoint setting of a bucket.

The up to date list of endpoints is provided by Amazon: https://docs.aws.amazon.com/general/latest/gr/rande.html#s3_region

Example

The following examples show per-bucket endpoints set for the "landsat-pds" and "eu-dataset" buckets, with the endpoints set to the default central endpoint and EU/Ireland, respectively:

```
<property>
  <name>fs.s3a.bucket.landsat-pds.endpoint</name>
  <value>s3.amazonaws.com</value>
  <description>The endpoint for s3a://landsat-pds URLs</description>
</property>
<property>
  <name>fs.s3a.bucket.eu-dataset.endpoint</name>
  <value>s3-eu-west-1.amazonaws.com</value>
  <description>The endpoint for s3a://eu-dataset URLs</description>
</property>
```

Explicitly declaring a bucket bound to the central endpoint ensures that if the default endpoint is changed to a new region, data stored in US-east is still reachable.

Encrypting Data on S3

Amazon S3 supports a number of encryption mechanisms to better secure the data in S3.

- In [Server-Side Encryption \(SSE\)](#), the data is encrypted before it is saved to disk in S3, and decrypted when it is read. This encryption and decryption takes place in the S3 infrastructure, and is transparent to (authenticated) clients.

- In [Client-Side Encryption \(CSE\)](#), the data is encrypted and decrypted on the client, that is, inside the AWS S3 SDK. This mechanism is not supported in Hadoop due to incompatibilities with most applications. Specifically, the amount of decrypted data is often less than the file length, breaking all the code which assumes that the content of a file is the same size as that stated in directory listings.



Note: CDP *only* supports Server-Side Encryption (SSE) and *does not* support Client-Side Encryption (CSE).

For this server-side encryption to work, the S3 servers require secret keys to encrypt data, and the same secret keys to decrypt it. These keys can be managed in three ways:

- SSE-S3: By using Amazon S3-Managed Keys
- SSE-KMS: By using AWS Key Management Service
- SSE-C: By using customer-supplied keys

In general, the specific configuration mechanism can be set via the property `fs.s3a.server-side-encryption-algorithm` in `core-site.xml`. However, some encryption options require extra settings. Server Side encryption slightly slows down performance when reading data from S3.

It is possible to configure encryption for specific buckets and to mandate encryption for a specific S3 bucket.

Related Information

[Troubleshooting S3 and S3Guard](#)

SSE-S3: Amazon S3-Managed Encryption Keys

In SSE-S3, all keys and secrets are managed inside S3. This is the simplest encryption mechanism.

Enabling SSE-S3

To write S3-SSE encrypted files, the value of `fs.s3a.server-side-encryption-algorithm` must be set to that of the encryption mechanism used in `core-site.xml`; currently only AES256 is supported.

```
<property>
  <name>fs.s3a.server-side-encryption-algorithm</name>
  <value>AES256</value>
</property>
```

Once set, all new data will be uploaded encrypted. There is no need to set this property when downloading data — the data will be automatically decrypted when read using the Amazon S3-managed key.

To learn more, refer to [Protecting Data Using Server-Side Encryption with Amazon S3-Managed Encryption Keys \(SSE-S3\)](#) in AWS documentation.



Note: When encrypted files are renamed, they are de-encrypted and then re-encrypted with the encryption settings algorithm of the client application performing the rename. If a client with encryption disabled renames an encrypted file, the new file will be unencrypted.

SSE-KMS: Amazon S3-KMS Managed Encryption Keys

Amazon offers a pay-per-use key management service, [AWS KMS](#). This service can be used to encrypt data on S3 using keys which can be centrally managed and assigned to specific roles and IAM accounts.

The AWS KMS [can be used by S3 to encrypt uploaded data](#). When uploading data encrypted with SSE-KMS, the named key that was used to encrypt the data is retrieved from the KMS service, and used to encode the per-object secret which encrypts the uploaded data. To decode the data, the same key must be retrieved from KMS and used to unencrypt the per-object secret key, which is then used to decode the actual file.

KMS keys can be managed by an organization's administrators in AWS, including having access permissions assigned and removed from specific users, groups, and IAM roles. Only those "principals" with granted rights to a key may access it, hence only they may encrypt data with the key, and decrypt data encrypted with it. This allows KMS to be used to provide a cryptographically secure access control mechanism for data stores on S3.

**Note:**

AWS KMS service is not related to the Key Management Service built into Hadoop (Hadoop KMS). The Hadoop KMS primarily focuses on managing keys for HDFS Transparent Encryption. Similarly, HDFS encryption is unrelated to S3 data encryption.

Enabling SSE-KMS

To enable SSE-KMS, the property `fs.s3a.server-side-encryption-algorithm` must be set to SSE-KMS in `core-site.xml`.

```
<property>
  <name>fs.s3a.server-side-encryption-algorithm</name>
  <value>SSE-KMS</value>
</property>
```

The ID of the specific key used to encrypt the data should also be set in the property `fs.s3a.server-side-encryption.key`:

```
<property>
  <name>fs.s3a.server-side-encryption.key</name>
  <value>arn:aws:kms:us-west-2:360379543683:key/071a86ff-8881-4ba0-9230-95af6d01ca01</value>
</property>
```

If your account is set up with a default KMS key and `fs.s3a.server-side-encryption.key` is unset, the default key will be used.

Alternatively, organizations may define a default key in the Amazon KMS; if a default key is set, then it will be used whenever SSE-KMS encryption is chosen and the value of `fs.s3a.server-side-encryption.key` is empty.



Note: [AWS Key Management Service \(KMS\)](#) is pay-per-use, working with data encrypted via KMS keys incurs extra charges during data I/O.

To learn more, refer to [Protecting Data Using Server-Side Encryption with AWS KMS-Managed Keys \(SSE-KMS\)](#) in the AWS documentation.



Note: The AWS KMS key used must be in the same region as the S3 Bucket where the data is being encrypted.



Note: The KMS key specified is only used when writing new data, or when renaming objects. When reading an existing object, whichever KMS key was used to encrypt the data is reused to decrypt it.

IAM Role permissions for working with SSE-KMS

All IAM roles which need to read data encrypted with SSE-KMS must have the permissions to decrypt using the specific key the data was encrypted with: `kms:Decrypt`

All IAM roles which need to both read and write data need the encrypt and decrypt permissions (encrypt-only permission is not supported).

```
kms:Decrypt
kms:GenerateDatakey
```

If a role does not have the permissions to read data, it will fail with an `java.nio.AccessDeniedException`.



Note: The renaming files require the permission to decrypt the data, as it is decrypted and then re-encrypted as it is copied. See [AWS KMS API Permissions: Actions and Resources Reference](#) for more details on KMS permissions.

SSE-C: Server-Side Encryption with Customer-Provided Encryption Keys

In SSE-C, the client supplies the secret key needed to read and write data.

The same SSE-C key must be used on all clients reading or writing the data; this must be set client-side, in the hadoop configuration.

For hadoop applications to work reliably, SSE-C with a common key *must* be used exclusively within a bucket if it is to be used at all. This is the only way to ensure that path and directory listings do not fail with "Bad Request" errors.

Enabling SSE-C

To use SSE-C, the configuration option `fs.s3a.server-side-encryption-algorithm` must be set to SSE-C, and a base-64 encoding of the key placed in `fs.s3a.server-side-encryption.key`.

```
<property>
  <name>fs.s3a.server-side-encryption-algorithm</name>
  <value>SSE-C</value>
</property>

<property>
  <name>fs.s3a.server-side-encryption.key</name>
  <value>RG8gbm90IGV2ZXIgbG9nIHRoaXMga2V5IG9yIG90aGVyd2lzMZSBzaGFyZSBpdA==</value>
</property>
```

This property can be set in a Hadoop JCEKS credential file, which is significantly more secure than embedding secrets in the XML configuration file.

Configuring Encryption for Specific Buckets

S3A's per-bucket configuration mechanism can be used to configure the encryption mechanism and credentials for specific buckets.

For example, to access the bucket called "production" using SSE-KMS with the key ID `arn:aws:kms:us-west-2:360379543683:key/071a86ff-8881-4ba0-9230-95af6d01ca01`, the settings are as follows:

```
<property>
  <name>fs.s3a.bucket.production.server-side-encryption-algorithm</name>
  <value>SSE-KMS</value>
</property>
```

```
<property>
  <name>fs.s3a.bucket.production.server-side-encryption.key</name>
  <value>arn:aws:kms:us-west-2:360379543683:key/071a86ff-8881-4ba0-9230-95af6d01ca01</value>
</property>
```

Encrypting an S3 Bucket with Amazon S3 Default Encryption

To guarantee that all data uploaded to a bucket is encrypted, it is possible to set a default encryption option for a bucket in the AWS management console.

For more information, see [Amazon S3 Default Encryption for S3 Buckets](#).

- This does not encrypt any data already stored in the bucket.
- S3A clients can still be configured to use a different encryption option if desired; this is the default value to use if no other policy was set.

A default encryption across a bucket offers significant benefits:

- It guarantees that all clients uploading data have encryption enabled.
- It guarantees that when a file is renamed, it will be re-encrypted, even if the client does not explicitly request encryption.
- If applied to an empty bucket, it guarantees that all future uploaded data in the bucket is encrypted.

We recommend selecting an encryption policy for a bucket when the bucket is created, and setting it in the bucket policy. This stops misconfigured clients from unintentionally uploading unencrypted data, or decrypting data when renaming files.

Performance Impact of Encryption

Server Side encryption slightly slows down performance when reading data from S3, both in the reading of data during the execution of a query, and in scanning the files prior to the actual scheduling of work.

Amazon [throttles reads and writes of S3-SSE data](#), which results in a significantly lower throughput than normal S3 IO requests. The default rate, 600 requests/minute, means that at most ten objects per second can be read or written using SSE-KMS per second — across an entire hadoop cluster (or the entire customer account). The default limits may be suitable during development — but in large scale production applications the limits may rapidly be reached. Contact Amazon to increase capacity.

Using S3Guard for Consistent S3 Metadata

S3Guard mitigates the issues related to S3's eventual consistency on listings by using a table on Amazon DynamoDB as a consistent metadata store. This guarantees a consistent view of data stored in S3.

In addition, S3Guard may improve query performance by reducing the number of times S3 needs to be contacted, — as DynamoDB is significantly faster.



Note: S3Guard should always be enabled on CDP in AWS as it is critical for guaranteeing the correctness of interactions with S3.

Introduction to S3Guard

Amazon S3 is an object store, not a filesystem. There are no directories, only objects. The S3A connector lets Hadoop, Hive and Spark applications see files in a directory tree, but really they are working on the objects underneath, by listing them and working on each one one-by-one.

Some of the operations which filesystems support are actually absent, with rename being the key one. The S3A connector mimics file or directory rename, by copying each file then deleting the original, which takes about 6-10 megabytes/second.

The S3 Object Store is "eventually consistent": when a file is deleted or overwritten it can take time for that change to propagate across all servers replicating the data. As a result, newly deleted files can still be visible, while queries of updated files can return the old version.

There is no specific time period by when the object store will be eventually consistent. The paper "[Benchmarking Eventual Consistency](#)" has shown it can vary by time of day, and be ten seconds or more — sometimes much more.

A critical problem is listing inconsistency: when a query is made of S3 to list all objects under a specific path, that listing can be out of date. This means that those operation on files under a "directory" mimicked by listing and acting on all objects underneath it are at risk of not seeing the complete list of files. Newly created files are at most risk.

This may affect the following operations on S3 data:

- When listing files, newly created objects may not be listed immediately and deleted objects may continue to be listed — which means that your input for data processing may be incorrect. In Hive, Spark, or MapReduce, this could lead to erroneous results. In the worst case, it could potentially lead to data loss at the time of data movement.
- When renaming directories, the listing may be incomplete or out of date, so the rename operation loses files. This is very dangerous as MapReduce, Hive, Spark and Tez all rely on rename to commit the output of workers to the final output of the job. If data is lost, the output is incorrect —something which may not be immediately obvious.
- When deleting directories, the listing may be inconsistent, so not all objects are deleted. If another job writes data to the same directory path, the old data may still be present.

- During an ETL workflow, in a sequence of multiple jobs that form the workflow, the next job is launched soon after the previous job has been completed. Applications such as Oozie rely on marker files to trigger the subsequent workflows. Any delay in the visibility of these files can lead to delays in the subsequent workflows.
- During existence-guarded path operations, if a deleted file which has the same name as a target path appears in a listing, some actions may unexpectedly fail due to the target path being present — even though the file has already been deleted.

This eventually consistent behavior of S3 can cause seemingly unpredictable results from queries made against it, limiting the practical utility of the S3A connector for use cases where data gets modified.

S3Guard mitigates the issues related to S3's eventual consistency on listings by using a table on Amazon DynamoDB as a consistent metadata store. This guarantees a consistent view of data stored in S3. In addition, S3Guard may improve query performance by reducing the number of times S3 needs to be contacted, —as DynamoDB is significantly faster.

How S3Guard Works

S3Guard is a feature in the Hadoop S3A connector which uses Amazon's DynamoDB to cache information about created and deleted files, "The S3Guard Database".

When an application using the S3A Connector with S3Guard enabled manipulates objects in S3, such as creating, deleting or "renaming" them, the S3Guard database is updated. Newly created/copied files are added to the table, while deleted files have "tombstone markers" added to indicate that they have been deleted.

When a directory is listed on S3, the information from the S3Guard database is used to update the listing with this information: new files are added while those files with tombstone markers are not included in a listing.

As a result, the listing is up to date with all operations performed on the directory by all clients using S3Guard.

When a file is opened for reading, or any existence check made, the S3Guard database is checked first. If an entry is found in the database, that is used as the response —omitting all checks of S3 itself. This includes tombstone markers, which are used as evidence that a file does not exist. The caller is given the current state of the object (existence/nonexistence, size and type), without S3 being queried at all. This can be significantly faster than interacting with S3.

With this design, directory listings are kept consistent across a sequence of operations, even across multiple servers —indeed, across multiple Hadoop clusters.

What S3Guard Cannot Do

S3Guard only stores filenames, lengths, "etags" and optional version data in its DynamoDB tables. The contents of the files are still stored in S3. If a file is overwritten, S3's eventual consistency may mean that the old data is still present. S3 tries to detect this by using the following techniques:

- Storing the "etag" checksum of each file and detecting when the value is different from that in S3.
- In an S3 bucket with "versioning" enabled, the version is used to guarantee that the file loaded is the one matching the S3Guard record.

When a file is loaded, the S3A connector will make repeated attempts to load a file with the expected etag and version ID before concluding that it is not available – at which point the operation will fail.

S3Guard does not provide the "atomic directory rename" operation which many applications rely on to commit their intermediate work to the final destination of the job. This does not work in S3 where files are copied one-by-one in a slow process. If this operation is interrupted, the source and destination will be in an unknown state.

- For Spark applications, use an S3A committer.
- When copying files using distcp, avoid the -atomic option. All it will do is make the copy slower. Use the -direct operation for a direct upload, and be aware that a failure during the copy can leave the destination inconsistent.

Configuring S3Guard

You can enable S3Guard when configuring your CDP environment.

For more information, see [Register an AWS environment](#) in the *Management Console* documentation.

Monitoring and Maintaining S3Guard

The DynamoDB console on AWS allows you to monitor the workload on DynamoDB.

If you are not using auto scaling, the AWS console allows you to adjust the read/write capacities of the table. The command `hadoop s3guard set-capacity` can also be used to alter these values.

Disabling S3Guard and destroying a table

To disable S3Guard, you must delete the per-bucket `fs.s3a.metastore.impl` parameter or set it back to the default `org.apache.hadoop.fs.s3a.s3guard.NullMetadataStore`.

If you are not using a shared table, in DynamoDB console on AWS, delete the DynamoDB table to avoid incurring unnecessary costs.

Destroying S3Guard table

The command `hadoop s3guard destroy` can be used on the command line to destroy a table.

```
hadoop s3guard destroy s3a://guarded-table/
2018-05-31 15:35:39,075 [main] INFO  s3guard.S3GuardTool (S3GuardTool.java:
initMetadataStore(270)) -
    Metadata store DynamoDBMetadataStore{region=eu-west-1, tableName
=guarded-table} is initialized.
2018-05-31 15:35:39,077 [main] INFO  s3guard.DynamoDBMetadataStore (DynamoDB
MetadataStore.java:destroy(793)) -
    Deleting DynamoDB table guarded-table in region eu-west-1
Metadata store is deleted.
```

Destroying a table does not destroy the data in S3; it merely removes the summary data used by S3Guard to provide consistent listings. You can verify this by listing the bucket:

```
> hadoop fs -ls -R s3a://guarded-table/
-rw-rw-rw-  1 alice alice          0 2019-07-31 20:45 s3a://guarded-table/
example
```



Note:

Even though `hadoop fs` commands can print the owner and permissions of a file, these are not real attributes of the object, and generated for the benefit of applications which check them. The owner and group of a file or directory is always the user for whom the S3A Filesystem was instantiated. File permissions are always read and write; while directories are read, write and execute — irrespective of any underlying access permissions on the objects.

Pruning Old Data from S3Guard Tables

S3Guard keeps tombstone markers of deleted files. It is good to clean these regularly, just to keep costs down. This can be done with the `hadoop s3guard prune` command.

The command can be used to delete entries older than a certain number of days, minutes or hours.

```
> hadoop s3guard prune -days 3 s3a://guarded-table/
2019-07-31 21:27:37,264 [main] INFO  s3guard.S3GuardTool (S3GuardTool.java:in
itMetadataStore(321)) - Metadata store DynamoDBMetadataStore{region=eu-west-
2, tableName=guarded-table, tableArn=arn:aws:dynamodb:eu-west-2:980678866538
:table/guarded-table} is initialized.
2019-07-31 21:27:37,285 [main] INFO  s3guard.DynamoDBMetadataStore (Duratio
nInfo.java:<init>(72)) - Starting: Pruning DynamoDB Store
2019-07-31 21:27:37,332 [main] INFO  s3guard.DynamoDBMetadataStore (Duratio
nInfo.java:close(87)) - Pruning DynamoDB Store: duration 0:00.047s
2019-07-31 21:27:37,332 [main] INFO  s3guard.DynamoDBMetadataStore (Dynamo
DBMetadataStore.java:innerPrune(1576)) - Finished pruning 366 items in batch
es of 25
```


Importing a Bucket into S3Guard

The `hadoop s3guard import` command can list and import a bucket's metadata into a S3Guard table. This is harmless if the contents are already imported.

```
hadoop s3guard import s3a://guarded-table/

2019-07-31 21:32:54,600 [main] INFO s3guard.S3GuardTool (S3GuardTool.java:in
itMetadataStore(321)) - Metadata store DynamoDBMetadataStore{region=eu-west-
2, tableName=guarded-table, tableArn=arn:aws:dynamodb:eu-west-2:980678866538
:table/guarded-table} is initialized.
Inserted 2 items into Metadata Store
```

You do not need to issue this command after creating a table; the data is added as listings of S3 paths find new entries. It merely saves by proactively building up the database.

Verifying that S3Guard is Enabled on a Bucket

When S3Guard is working, apart from some messages in the logs, there is no obvious clue that it is enabled. To verify that a bucket does have S3Guard enabled, use the command-line command `hadoop s3guard bucket-info`. This will print bucket information, and can be used to explicitly check that a bucket has s3guard enabled.

On a guarded bucket, it will list details about the bucket and the S3Guard Database on DynamoDB.

```
hadoop s3guard bucket-info s3a://example1/
Filesystem s3a://example1
Location: eu-west-1
Filesystem s3a://example1 is using S3Guard with store DynamoDBMetadataStore
{region=eu-west-1, tableName=example1, tableArn=arn:aws:dynamodb:eu-west-1:9
8067886600:table/example1}
Authoritative Metadata Store: fs.s3a.metadatastore.authoritative=false
Authoritative Path: fs.s3a.authoritative.path=
Metadata Store Diagnostics:
  ARN=arn:aws:dynamodb:eu-west-1:98067886600:table/example1
  billing-mode=per-request
  description=S3Guard metadata store in DynamoDB
  name=example1
  persist.authoritative.bit=true
  read-capacity=0
  region=eu-west-1
  retryPolicy=ExponentialBackoffRetry(maxRetries=9, sleepTime=250 MILLISECOND
S)
  size=14431
  status=ACTIVE
  table={AttributeDefinitions: [{AttributeName: child,AttributeType: S}, {A
ttributeName: parent,AttributeType: S}],TableName: example1,KeySchema: [{Att
ributeName: parent,KeyType: HASH}, {AttributeName: child,KeyType: RANGE}],Ta
bleStatus: ACTIVE,CreationDateTime: Wed Jun 05 13:23:18 BST 2019,Provisioned
Throughput: {NumberOfDecreasesToday: 0,ReadCapacityUnits: 0,WriteCapacityUni
ts: 0},TableSizeBytes: 14431,ItemCount: 96,TableArn: arn:aws:dynamodb:eu-wes
t-1:98067886600:table/example1,TableId: b4b1b660-9cbd-0000f10623,BillingMode
Summary: {BillingMode: PAY_PER_REQUEST,LastUpdateToPayPerRequestDateTime: We
d Jun 05 13:32:07 BST 2019}},}
  write-capacity=0
The "magic" committer is supported

S3A Client
Signing Algorithm: fs.s3a.signing-algorithm=(unset)
Endpoint: fs.s3a.endpoint=s3-eu-west-1.amazonaws.com
Encryption: fs.s3a.server-side-encryption-algorithm=none
Input seek policy: fs.s3a.experimental.input.fadvise=normal
Change Detection Source: fs.s3a.change.detection.source=etag
Change Detection Mode: fs.s3a.change.detection.mode=server
```

```
Delegation token support is disabled
```

In this example the bucket is shown to be guarded by a table whose billing mode is "per-request".

Using the S3Guard CLI

The S3Guard CLI offers other maintenance commands. for information on how to use them, refer to [Apache documentation](#).

One useful command to use is `hadoop s3guard prune -tombstone`, which removes "tombstone markers" from the table. These are entries created when files are deleted, so as to identify recently deleted files which should be omitted from listings. After the S3 Store has become consistent with these deletions, the markers are no longer deleted.

```
> hadoop s3guard prune -tombstone -days 1 s3a://hwdev-steve-ireland-new/

2019-08-20 16:55:40,790 [main] INFO  s3guard.S3GuardTool - Metadata store
DynamoDBMetadataStore{region=eu-west-1, tableName=example1, tableArn=arn:aws:
:dynamodb:eu-west-1:98067886600:table/example1} is initialized.
2019-08-20 16:55:40,810 [main] INFO  s3guard.DynamoDBMetadataStore - Startin
g: Pruning DynamoDB Store
2019-08-20 16:55:40,846 [main] INFO  s3guard.DynamoDBMetadataStore - Pruning
DynamoDB Store: duration 0:00.036s
2019-08-20 16:55:40,846 [main] INFO  s3guard.DynamoDBMetadataStore - Finis
hed pruning 0 items in batches of 25
```

S3Guard: Operational Issues

There are various issues and limitations that you must consider when working with S3Guard.

The following operational issues have been identified while testing S3Guard.

Third-party S3-compatible object stores

Third-party object stores which reimplement the AWS S3 protocol are usually "consistent". As such, there is no need to use S3Guard. Consult the object store's supplier as to its consistency model.

S3Guard Security Aspects

The DynamoDB table needs to be writeable by all users/services using S3Guard. If a single DynamoDB table is used to store metadata about multiple buckets, then clients with access to the table will be able to read the metadata about objects in any bucket to which their read access restricted via AWS permissions.

The standard S3 Bucket and Object Access permissions do not provide any restriction on accessing the S3Guard index data. As this is only the Hadoop file status data of object name, type, size and timestamp, the actual object data and any tags attached to the object are still protected by AWS permissions. However, directory and filenames will be visible.

Limitations of S3Guard

The key limitation of S3Guard is that it only provides consistent file and directory listings. It does not address update and delete consistency of the data. It is only consistent with respect to changes made by client applications using the S3A connector with S3Guard enabled and the same DynamoDB table. Changes which are made by other applications are only eventually consistent from the perspective of S3A clients.

S3Guard will track the etag of uploaded files, and, on a versioned S3 bucket, the S3 version ID of an uploaded file. These will be used when opening files, so as to detect and possibly react to changes.

Safely Writing to S3 Through the S3A Committers

The S3A committers are three different committers used to commit work directly to Mapreduce and Spark. The committers are enabled by default for Spark in CDP.

Introducing the S3A Committers

Amazon's S3 Object Store is not a filesystem: some expected behaviors of a filesystem are missing.

Some of the following aspects of S3 that are different from a filesystem are as follows:

- Directory listings are only eventually consistent.
- File overwrites and deletes are only eventually consistent: readers may get old data.
- There is no rename operation; it is mimicked in the S3A client by listing a directory and copying each file underneath, one-by-one.

Because directory rename is mimicked by listing and then copying files the eventual consistency of both listing and reading fails may result in incorrect data. And, because of the copying: it is slow.

S3Guard addresses the listing inconsistency problem. However, it does not address the update consistency or performance.

The normal means by which Hadoop MapReduce and Apache Spark commit work from multiple tasks is through renaming the output. Each task attempt writes to a private task attempt directory; when the task is given permission to commit by the MapReduce Application Master or Spark Driver, this task attempt directory is renamed into the job attempt directory. When the job is ready to commit, the output of all the tasks is merged into the final output directory, again by renaming files and directories.

This is fast and safe on HDFS and similar filesystems, and on object stores with rename operations, such as Azure WASB. On S3, it is unreliable without S3Guard, and even with S3Guard, the time to commit work to S3 is proportional to the amount of data written. An operation which may work well during development can turn out to be unusable in production.

To address this the S3A committers were developed. They allow the output of MapReduce and Spark jobs to be written directly to S3, with a time to commit the job independent of the amount of data created.

What Are the S3A Committers?

The S3A committers are three different committers which can be used to commit work directly to Map-reduce and Spark. They differ in how they deal with conflict and how they upload data to the destination bucket—but underneath they all share much of the same code.

They rely on a specific S3 feature: multipart upload of large files.

When writing a large object to S3, S3A and other S3 clients use a mechanism called “Multipart Upload”.

The caller initiates a “multipart upload request”, listing the destination path and receiving an upload ID to use in the upload operations.

The caller then uploads the data in a series of HTTP POST requests, closing with a final POST listing the blocks uploaded.

The uploaded data is not visible until that final POST request is made, at which point it is published in a single atomic operation.

This mechanism is always used by S3A whenever it writes large files; the size of each part is set to the value of `fs.s3a.multipart.size`

The S3A Committers use the same multipart upload process, but convert it into a mechanism for committing the work of tasks because of a special feature of the mechanism: The final POST request does not need to be issued by the same process or host which uploaded the data.

The output of each worker task can be uploaded to S3 as a multipart upload, but without issuing the final POST request to complete the upload. Instead, all the information needed to issue that request is saved to a cluster-wide filesystem (HDFS or potentially S3 itself)

When a job is committed, this information is loaded, and the upload completed. If a task is aborted or fails: the upload is not completed, so the output does not become visible. If the entire job fails, none of its output becomes visible.

For further reading, see:

- [HADOOP-13786](#): Add S3A committers for zero-rename commits to S3 endpoints.
- [S3A Committers: Architecture and Implementation](#).
- [A Zero-Rename Committer](#): Object-storage as a Destination for Apache Hadoop and Spark.

The Three Committers

The three different S3A committers are directory committer, partitioned committer, magic committer.

- **Directory Committer**: Buffers working data to the local disk, uses HDFS to propagate commit information from tasks to job committer, and manages conflict across the entire destination directory tree.
- **Partitioned Committer**: Identical to the Directory committer except that conflict is managed on a partition-by-partition basis. This allows it to be used for in-place updates of existing datasets. It is only suitable for use with Spark.
- **Magic Committer**: Data is written directly to S3, but “magically” re-targeted at the final destination. Conflict is managed across the directory tree. It requires a consistent S3 object store, which means S3Guard is a mandatory pre-requisite.

We currently recommend use of the “Directory” committer: it is the simplest of the set, and by using HDFS to propagate data from workers to the job committer, does not directly require S3Guard – this makes it easier to set up.

The rest of the documentation only covers the Directory Committer: to explore the other options, consult the Apache documentation.

Configuring Directories for Intermediate Data

In addition to `fs.s3a.committer.name`, two other `core-site.xml` configuration options are used to control where intermediate is stored.

A location is in the local filesystem for buffering data

```
<property>
  <name>fs.s3a.buffer.dir</name>
  <value>${hadoop.tmp.dir}/s3a</value>
  <description>Comma separated list of directories that will be used to buffer file uploads to.</description>
</property>
```

These directories will store the output created by all active tasks until each task is committed; the more worker processes/spark worker threads a host can support, the more disk space will be needed. Multiple disks can be listed to help spread the load, and recover from disk failure.

A location in the cluster's HDFS filesystem to share summary data about pending uploads.

```
<property>
  <name>fs.s3a.committer.staging.tmp.path</name>
  <value>tmp/staging</value>
</property>
```

These files are generally quite small: a few kilobytes per task.

Using the Directory Committer in MapReduce

Once the property `fs.s3a.committer.name` is set, Hadoop MapReduce jobs writing to paths using the `s3a://` schema will automatically switch to the new committers.

Jobs using any other filesystem as a destination will still use the normal file committer.

Verifying That an S3A Committer Was Used

When working correctly, the only sign the new committers are in use is that it should be faster to use S3 as a destination of work.

There is a straightforward way to determine if a new committer was used: examine the `_SUCCESS` file created in the destination directory of a query. With the original file committer, this is a zero-byte file. The new S3A committers all write a JSON file describing the committer used, the files created and various diagnostics information.

Listing this file is enough to show whether an S3A committer was used:

```
hadoop fs -ls s3a://guarded-bucket/datasets/orc/_SUCCESS
```

If this file is of size 0, then no S3A committer was used. If the file length is greater than zero, then an S3A committer was used.

To see more details about the job commit operation, the file's contents can be printed.

```
hadoop fs -cat s3a://guarded-bucket/datasets/orc/_SUCCESS
```

Cleaning up after failed jobs

The S3A committers upload data in the tasks, completing the uploads when the job is committed.

About this task

Amazon AWS still bill you for all data held in this “pending” state. The `hadoop s3guard uploads` command can be used to list and cancel such uploads. However, it is simplest to automate cleanup with a bucket lifecycle rule.

Procedure

1. Go to the AWS S3 console: <https://s3.console.aws.amazon.com/>.
2. Find the bucket you are using as a destination of work.
3. Select the Management tab.
4. Select Add a new lifecycle rule.
5. Create a rule “cleanup uploads” with no filter, and without any “transitions”.
Configure an “Expiration” action of Clean up incomplete multipart uploads.
6. Select a time limit for outstanding uploads, such as 1 Day.
7. Review and confirm the lifecycle rule

You need to select a limit of how long uploads can be outstanding. For Hadoop applications, this is the maximum time that either an application can write to the same file and the maximum time which a job may take. If the timeout is shorter than either of these, then programs are likely to fail.

Once the rule is set, the cleanup is automatic.

Using the S3Guard Command to List and Delete Uploads

The `hadoop s3guard uploads` command can also be used to list all outstanding uploads under a path, and delete them.

```
hadoop s3guard uploads s3a://guarded-bucket/
tests3ascale/scale/hugefile nB3gctgP34ZSpzJ5_o2AVb7kKRiicXfbkBqIf1A0y.IH7Ci
mH100VUzohRCj3QfFstoQIdcge478ZidXu764yN0vuGI1j5kCOV3rDhsrc.RBZ5skZ93jVCN9g2c
21QgB
```

```
Total 1 uploads found.
```

All outstanding uploads can be aborted, or those older than a certain age.

```
hadoop s3guard uploads -abort -days 1 -force s3a://guarded-bucket/
Deleting: tests3ascale/scale/hugefile nB3gctgP34ZSpzJ5_o2AVb7kKRiicXfbkBq
If1A0y.IH7CimH100VUzohRCj3QfFstoQIdcge478ZidXu764yN0vuGI1j5kcOV3rDhsrc.RBZ5s
kZ93jVCN9g2c21QgB
2018-06-28 20:58:18,504 [main] INFO s3a.S3AFileSystem (S3AFileSystem.java:
abortMultipartUpload(3266)) - Aborting multipart upload nB3gctgP34ZSpzJ5_o2A
Vb7kKRiicXfbkBqIf1A0y.IH7CimH100VUzohRCj3QfFstoQIdcge478ZidXu764yN0vuGI1j5kc
OV3rDhsrc.RBZ5skZ93jVCN9g2c21QgB to tests3ascale/scale/hugefile
Total 1 uploads deleted.
```

While bucket lifecycle is the best way to guarantee that all outstanding uploads are deleted; this command line tool is useful to verify that such cleanup is taking place, or explicitly clean up after a failure.

Advanced Committer Configuration

The Apache documentation covers the full set of configuration options for the committers.

Enabling Speculative Execution

The S3A committers all support speculative execution.

For MapReduce, enable the following properties in the job:

```
mapreduce.map.speculative true
mapreduce.reduce.speculative true
```

For Spark, set the following configuration option:

```
spark.speculation true
```

Using Unique Filenames to Avoid File Update Inconsistency

When updating existing datasets, if a new file overwrites an existing file of the same name, S3's eventual consistency on file updates means that the old data may still be returned.

To avoid this, unique filenames should be used when creating files. The property `fs.s3a.committer.staging.unique-file-names` enables this.

```
<property>
  <name>fs.s3a.committer.staging.unique-filename</name>
  <value>true</value>
</property>
```

It is set to true by default; you only need to disable it if you explicitly want newly created files to have the same name as any predecessors.

Speeding up Job Commits by Increasing the Number of Threads

When an individual job writes many files to S3, the time to commit the job can increase.

It can be speeded up by increasing the value of `fs.s3a.committer.threads`. However, the value of `fs.s3a.connection.maximum` must be at least this size otherwise the limit on the number of parallel uploads is still limited.

```
<property>
  <name>fs.s3a.committer.threads</name>
  <value>8</value>
</property>
<property>
  <name>fs.s3a.connection.maximum</name>
  <value>15</value>
```

```
</property>
```

Securing the S3A Committers

There are a few security considerations when using the S3A committers.

S3 Bucket Permissions

To use an S3A committer, the account/role must have the same AWS permissions as needed to write to the destination bucket.

The multipart upload list/abort operations may be a new addition to the permissions for the active role.

When using S3Guard, the account/role must also have full read/write/delete permissions for the DynamoDB table used.

Local Filesystem Security

All the committers use the directories listed in `fs.s3a.buffer.dir` to store staged/buffered output before uploading it to S3.

To ensure that other processes running on the same host do not have access to this data, unique paths should be used per-account.

This requirement holds for all applications working with S3 through the S3A connector.

HDFS Security

The directory and partitioned committers use HDFS to propagate commit information between workers and the job committer.

These intermediate files are saved into a job-specific directory under the path `${fs.s3a.committer.staging.tmp.path}/${user}` where `${user}` is the name of the user running the job.

The default value of `fs.s3a.committer.staging.tmp.path` is `tmp/staging`, Which will be converted at run time to a path under the current user's home directory, essentially `/user/${user}/tmp/staging/${user}/`.

Provided the user's home directory has access restricted to trusted accounts, this intermediate data will be secure.

If the property `fs.s3a.committer.staging.tmp.path` is changed to a different location, then this path will need to be secured to protect pending jobs from tampering.



Note: This intermediate data does not contain the output, merely the listings of all pending files and the MD5 checksums of each block.

The S3A Committers and Third-Party Object Stores

The S3A committers will work with any object store which implements the AWS S3 protocols.

The directory committer requires a consistent cluster filesystem to propagate commit information from the worker processes to the job committer. This is normally done in HDFS.

If the third-party object store is consistent, then it may also be used as the cluster filesystem. Set `fs.s3a.committer.staging.tmp.path` to a chosen path in the store.

Limitations of the S3A Committers

There are limitations of the S3A committers associated with custom file output formats, MapReduce API output format, and non-availability of Hive support.

Custom file output formats and their committers

Output formats which implement their own committers do not automatically switch to the new committers. If such a custom committer relies on renaming files to commit output, then it will depend on S3Guard for a consistent view of the object store, and take time to commit output proportional to the amount of data and the number of files.

To determine if this is the case, find the subclass of `org.apache.hadoop.mapreduce.lib.output.FileOutputFormat` which implements the custom format, to see if it subclasses the `getOutputCommitter()` to return its own committer, or has a custom subclass of `org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter`.

It may be possible to migrate such a committer to support store-specific committers, as was done for Apache Parquet support in Spark. Here a subclass of Parquet's `ParquetOutputCommitter` was implemented to delegates all operations to the real committer.

MapReduce V1 API Output Format and Committers

Only the MapReduce V2 APIs under `org.apache.hadoop.mapreduce` support the new committer binding mechanism. The V1 APIs under the `org.apache.hadoop.mapred` package only bind to the file committer and subclasses. The v1 APIs date from Hadoop 1.0 and should be considered obsolete. Please migrate to the v2 APIs, not just for the new committers, but because the V2 APIs are still being actively developed and maintained.

No Hive Support

There is currently no Hive support for the S3A committers. To safely use S3 as a destination of Hive work, you must use `S3Guard`.

Troubleshooting the S3A Committers

The [Apache documentation](#) contains information about troubleshooting the committers.

The primary issue which surfaces is actually programs not switching to the new committers. There are three common reasons for this.

- The configuration settings to switch to the new committer are not being picked up. This is particularly common in Spark, which is harder to set up.
- The program is using the older V1 MapReduce APIs. Fix: switch to the V2 API.
- The output format the program uses is explicitly creating its own committer. This can only be fixed by modifying the program.

To help debug Spark's configuration, there is an option which can be set to forcibly fail the spark query if the path output committer is used. However, the file committer is being returned.

```
spark.hadoop.pathoutputcommit.reject.fileoutput true
```

There is also the blunt-instrument approach of causing the original output committer to crash with an invalid configuration.

```
spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version 3
```

This (invalid) option ensures that if the original file committer is used, it will raise an exception.

To enable low-level logging of the committers, set the log-level of the package `org.apache.hadoop.fs.s3a.commit` to `DEBUG`. With `Log4J`, this can be one in `log4j.properties`:

```
log4j.logger.org.apache.hadoop.fs.s3a.commit=DEBUG
```

Security Model and Operations on S3

The security and permissions model of Amazon S3 is very different from this of a UNIX-style filesystem: on Amazon S3, operations which query or manipulate permissions are generally unsupported. Operations to which this applies include: `chgrp`, `chmod`, `chown`, `getfacl`, and `setfacl`. The related attribute commands `getfattr` and `setfattr` are also unavailable.

In addition, operations which try to preserve permissions (for example `fs -put -p`) do not preserve permissions.

Although these operations are unsupported, filesystem commands which list permission and user/group details usually simulate these details. As a consequence, when interacting with read-only object stores, the permissions found in “list” and “stat” commands may indicate that the user has write access — when they do not.

Amazon S3 has a permissions model of its own. This model can be manipulated through store-specific tooling. Be aware that some of the permissions which can be set — such as write-only paths, or various permissions on the root path — may be incompatible with the S3A client. It expects full read and write access to the entire bucket with trying to write data, and may fail if it does not have these permissions.

As an example of how permissions are simulated, here is a listing of Amazon's public, read-only bucket of Landsat images:

```
$ hadoop fs -ls s3a://landsat-pds/
Found 10 items
drwxrwxrwx - mapred 0 2016-09-26 12:16 s3a://landsat-pds/L8
-rw-rw-rw- 1 mapred 23764 2015-01-28 18:13 s3a://landsat-pds/index.html
drwxrwxrwx - mapred 0 2016-09-26 12:16 s3a://landsat-pds/landsat-pds_stats
-rw-rw-rw- 1 mapred 105 2016-08-19 18:12 s3a://landsat-pds/robots.txt
-rw-rw-rw- 1 mapred 38 2016-09-26 12:16 s3a://landsat-pds/run_info.json
drwxrwxrwx - mapred 0 2016-09-26 12:16 s3a://landsat-pds/runs
-rw-rw-rw- 1 mapred 27458808 2016-09-26 12:16 s3a://landsat-pds/scene_list.
gz
drwxrwxrwx - mapred 0 2016-09-26 12:16 s3a://landsat-pds/tarq
drwxrwxrwx - mapred 0 2016-09-26 12:16 s3a://landsat-pds/tarq_corrupt
drwxrwxrwx - mapred 0 2016-09-26 12:16 s3a://landsat-pds/test
```

The following is evident from the examples:

- All files are listed as having full read/write permissions.
- All directories appear to have full rwx permissions.
- The replication count of all files is "1".
- The owner of all files and directories is declared to be the current user (mapred).
- The timestamp of all directories is actually that of the time the `-ls` operation was executed. This is because these directories are not actual objects in the store; they are simulated directories based on the existence of objects under their paths.

When an attempt is made to delete one of the files, the operation fails — despite the permissions shown by the `ls` command:

```
$ hadoop fs -rm s3a://landsat-pds/scene_list.gz
rm: s3a://landsat-pds/scene_list.gz: delete on s3a://landsat-pds/scene_list.
gz:
com.amazonaws.services.s3.model.AmazonS3Exception: Access Denied (Service: A
mazon S3;
Status Code: 403; Error Code: AccessDenied; Request ID: 1EF98D5957BCAB3D),
S3 Extended Request ID: wi3veOXFuFqWBUCJgV3Z+NQVj9gWgZVdXlPU4KBbYMsw/gA+hyh
RXcaQ+PogOsDgHh3lHlTCebQ=
```

This demonstrates that the listed permissions cannot be taken as evidence of write access. Only object manipulation can determine this.

S3A and Checksums (Advanced Feature)

The S3A connector can be configured to export the HTTP etag of an object as a checksum, by setting the option `fs.s3a.etag.checksum.enabled` to `true`. When unset (the default), S3A objects have no checksum.

```
$ hadoop fs -touchz s3a://hwdev-bucket/src/something.txt
$ hadoop fs -checksum s3a://hwdev-bucket/src/something.txt
```

```
s3a://hwdev-bucket/src/something.txt NONE
```

Once set, S3A objects have a checksum which is created on upload.

```
$ hadoop fs -Dfs.s3a.etag.checksum.enabled=true -checksum s3a://hwdev-bucket/
src/something.txt
s3a://hwdev-bucket/src/something.txt etag 6434316438636439386630306232303
465393830303939386563663834323765
```

This checksum is not compatible with that of HDFS, so cannot be used to compare file versions when using the `-update` option on DistCp between S3 and HDFS. More specifically, unless `-skipcrccheck` is set, the DistCP operation will fail with a checksum mismatch. However, it can be used for incremental updates within and across S3A buckets.

```
$ hadoop distcp -Dfs.s3a.etag.checksum.enabled=true --update s3a://hwdev-buc
ket/src s3a://hwdev-bucket/dest
$ hadoop fs -Dfs.s3a.etag.checksum.enabled=true -checksum s3a://hwdev-bucke
t/dest/something.txt
s3a://hwdev-bucket/src/something.txt etag 643431643863643938663030623230346
5393830303939386563663834323765
```

As the checksums match small files created as a single block, incremental updates will not copy unchanged files. For large files uploaded using multiple blocks, the checksum values may differ in which case the source file will be copied again.

A List of S3A Configuration Properties

All S3A client options are configured with options with the prefix `fs.s3a`.

For information about the S3A configuration properties, see [General S3A client configuration](#).

Working with versioned S3 buckets

AWS S3 supports "Versioning" in buckets, where the bucket is configured to save older versions of objects. When an object is overwritten (or even deleted), the old version can be accessed when a request is made for the object using the original version ID.

For more information, see [Using Versioning](#).

The S3A connector supports versioning in the following ways:

- S3Guard will record the `versionId` of new files written into the store. It can then use that value when opening a file, so that it can guarantee that the specific version listed in the S3Guard table is opened. This can compensate for eventual consistency of overwritten data—even if the old version is initially found when opening the file, the S3A connector can retry until the new version is found.
- When a file is opened for reading, the version ID of that file is recorded, and then for the duration of the file read (seconds, minutes, hours...) only that version of the data is read. Even if the file is overwritten, the single ongoing file read will always read the original data.
- When a file is copied (as in a rename operation), the version ID is used to guarantee that even if the source file is overwritten, the copied file will be the original version.
- It can be used as an alternative to moving deleted files to a trash location: simply delete the files and then recover them later. Note: The S3A connector does not provide a recovery tool.

The following are some issues you must be aware of when using versioning:

- Too many S3 Tombstone markers from deleted objects will slow down directory listings, and can result in clients being throttled (<https://docs.aws.amazon.com/AmazonS3/latest/dev/troubleshooting.html#troubleshooting-by-symptom-increase-503-reponses>)

- Old versions of files are still billed for.

To keep costs down and minimize performance problems, we recommend having a lifecycle rule on the bucket which deletes older versions of files after a number of days. This can be done from the Management tab of the AWS S3 console.

Here is an example policy which deletes all versions of objects which were overwritten more than a week previously, while also cleaning up tombstone markers and incomplete file uploads:

Lifecycle rule

1 Name and scope 2 Transitions **3 Expiration** 4 Review

Configure expiration

☐ Current version ☒ Previous versions

☒ Permanently delete previous versions ⓘ

After days from becoming a previous version

Clean up expired object delete markers and incomplete multipart uploads

☒ Clean up expired object delete markers ⓘ

☒ Clean up incomplete multipart uploads ⓘ

After days from start of upload

[Previous](#) [Next](#)

Working with Third-party S3-compatible Object Stores

The S3A Connector can work with third-party object stores; some vendors test the connector against their stores — and even actively collaborate in developing the connector in the open source community.

Option	Change to
fs.s3a.endpoint	(the hostname of the S3 Store)
fs.s3a.path.style.access	true

Option	Change to
fs.s3a.signing-algorithm	If the default signing mechanism is rejected, another mechanism may work from: "QueryStringSignerType", "S3SignerType", "AWS3SignerType", "AWS4SignerType", "AWS4UnsignedPayloadSignerType" and "NoOpSignerType".
fs.s3a.connection.ssl.enabled	Set to "false" if HTTP is used instead of HTTPS
fs.s3a.multiobjectdelete.enable	Set to "false" if bulk delete operations are not supported.
fs.s3a.list.version	Set to "1" if the list directories with the default "2" option fails with an error.

Third-party object stores are generally consistent; there is no need for S3Guard. The S3A Committers will still offer better performance, and should be used for MapReduce and Spark.

Encryption may or may not be supported: consult the documentation.

Security permissions are likely to be implemented differently from the IAM role mode —again, consult the documentation to see what is available.

Improving Performance for S3A

You can consider various options for improving performance when working with data stored in Amazon S3.

The bandwidth between the CDP cluster and Amazon S3 is the upper limit to how fast data can be copied into S3. The further the CDP cluster is from the Amazon S3 installation, or the narrower the network connection is, the longer the operation will take. Even a CDP cluster deployed within Amazon's own infrastructure may encounter network delays from throttled VM network connections.

Network bandwidth limits notwithstanding, there are some options which can be used to tune the performance of an upload:

- [Working with S3 buckets in the same AWS region](#) on page 28
- [Configuring and tuning S3A block upload](#) on page 28

Working with S3 buckets in the same AWS region

A foundational step to getting good performance is working with buckets close to the Hadoop cluster, where "close" is measured in network terms.

Maximum performance is achieved from working with S3 buckets in the same AWS region as the cluster. For example, if your cluster is in North Virginia ("US East"), you will achieve best performance if your S3 bucket is in the same region.

In addition to improving performance, working with local buckets ensures that no bills are incurred for reading from the bucket.

Configuring and tuning S3A block upload

Because of the nature of the S3 object store, data written to an S3A OutputStream is not written incrementally — instead, by default, it is buffered to disk until the stream is closed in its close() method. This can make output slow because the execution time for OutputStream.close() is proportional to the amount of data buffered and inversely proportional to the bandwidth between the host to S3; that is $O(\text{data}/\text{bandwidth})$.

Other work in the same process, server, or network at the time of upload may increase the upload time.

In summary, the further the process is from the S3 store, or the smaller the EC2 VM is, the longer it will take complete the work. This can create problems in application code:

- Code often assumes that the close() call is fast; the delays can create bottlenecks in operations.
- Very slow uploads sometimes cause applications to time out - generally, threads blocking during the upload stop reporting progress, triggering timeouts.

- Streaming very large amounts of data may consume all disk space before the upload begins.

Tuning S3A Uploads

When data is written to S3, it is buffered locally and then uploaded in multi-Megabyte blocks, with the final upload taking place as the file is closed.

The following major configuration options are available for the S3A block upload options. These are used whenever data is written to S3.

Parameter	Default Value	Description
fs.s3a.multipart.size	100M	Defines the size (in bytes) of the blocks into which the upload or copy operations will be split up. A suffix from the set {K,M,G,T,P} may be used to scale the numeric value.
fs.s3a.fast.upload.active.blocks	8	Defines the maximum number of blocks a single output stream can have active uploading, or queued to the central FileSystem instance's pool of queued operations. This stops a single stream overloading the shared thread pool.
fs.s3a.buffer.dir	Empty value	A comma separated list of temporary directories use for storing blocks of data prior to their being uploaded to S3. When unset (by default), the Hadoop temporary directory <code>hadoop.tmp.dir</code> is used.
fs.s3a.fast.upload.buffer	disk	<p>The <code>fs.s3a.fast.upload.buffer</code> determines the buffering mechanism to use when uploading data.</p> <p>Allowed values are: disk, array, bytebuffer:</p> <ul style="list-style-type: none"> • (default) "disk" will use the directories listed in <code>fs.s3a.buffer.dir</code> as the location(s) to save data prior to being uploaded. • "array" uses arrays in the JVM heap. • "bytebuffer" uses off-heap memory within the JVM. <p>Both "array" and "bytebuffer" will consume memory in a single stream up to the number of blocks set by: <code>fs.s3a.multipart.size * fs.s3a.fast.upload.active.blocks</code>. If using either of these mechanisms, keep this value low.</p> <p>The total number of threads performing work across all threads is set by <code>fs.s3a.threads.max</code>, with <code>fs.s3a.max.total.tasks</code> values setting the number of queued work items.</p>

Note that:

- If the amount of data written to a stream is below that set in `fs.s3a.multipart.size`, the upload takes place after the application has written all its data.
- The maximum size of a single file in S3 is one thousand blocks, which, for uploads means $10000 * fs.s3a.multipart.size$. Too A small value of `fs.s3a.multipart.size` can limit the maximum size of files.
- Incremental writes are not visible; the object can only be listed or read when the multipart operation completes in the `close()` call, which will block until the upload is completed.

Buffering uploads to disk or RAMs

This is the default buffer mechanism. The amount of data which can be buffered is limited by the amount of available disk space.

When `fs.s3a.fast.upload.buffer` is set to "disk", all data is buffered to local hard disks prior to upload. This minimizes the amount of memory consumed, and so eliminates heap size as the limiting factor in queued uploads.

Buffering uploads in Byte Buffers

When `fs.s3a.fast.upload.buffer` is set to "bytebuffer", all data is buffered in "direct" ByteBuffers prior to upload. This may be faster than buffering to disk in cases such as when disk space is small there may not be much disk space to buffer with (for example, when using "tiny" EC2 VMs).

The ByteBuffers are created in the memory of the JVM, but not in the Java Heap itself. The amount of data which can be buffered is limited by the Java runtime, the operating system, and, for YARN applications, the amount of memory requested for each container.

The slower the upload bandwidth to S3, the greater the risk of running out of memory — and so the more care is needed in tuning the upload thread settings to reduce the maximum amount of data which can be buffered awaiting upload.

Buffering Uploads with Array Buffers

When `fs.s3a.fast.upload.buffer` is set to "array", all data is buffered in byte arrays in the JVM's heap prior to upload. This may be faster than buffering to disk.

The amount of data which can be buffered is limited by the available size of the JVM heap. The slower the write bandwidth to S3, the greater the risk of heap overflows. This risk can be mitigated by tuning the upload thread settings.

Thread Tuning for S3A Data Upload

Both the array and bytebuffer buffer mechanisms can consume very large amounts of memory, on-heap or off-heap respectively. The disk buffer mechanism does not use much memory up, but it consumes hard disk capacity.

If there are many output streams being written to in a single process, the amount of memory or disk used is the multiple of all streams' active memory and disk use.

You may need to perform careful tuning to reduce the risk of running out memory, especially if the data is buffered in memory. There are a number parameters which can be tuned:

Parameter	Default Value	Description
<code>fs.s3a.fast.upload.active.blocks</code>	4	Maximum number of blocks a single output stream can have active (uploading, or queued to the central FileSystem instance's pool of queued operations). This stops a single stream overloading the shared thread pool.
<code>fs.s3a.threads.max</code>	10	The total number of threads available in the filesystem for data uploads or any other queued filesystem operation.
<code>fs.s3a.max.total.tasks</code>	5	The number of operations which can be queued for execution
<code>fs.s3a.threads.KeepAliveTime</code>	60	The number of seconds a thread can be idle before being terminated.

When the maximum allowed number of active blocks of a single stream is reached, no more blocks can be uploaded from that stream until one or more of those active block uploads completes. That is, a `write()` call which would trigger an upload of a now full datablock will instead block until there is capacity in the queue.

Consider the following:

- As the pool of threads set in `fs.s3a.threads.max` is shared (and intended to be used across all threads), a larger number here can allow for more parallel operations. However, as uploads require network bandwidth, adding more threads does not guarantee speedup.

- The extra queue of tasks for the thread pool (`fs.s3a.max.total.tasks`) covers all ongoing background S3A operations.
- When using memory buffering, a small value of `fs.s3a.fast.upload.active.blocks` limits the amount of memory which can be consumed per stream.
- When using disk buffering, a larger value of `fs.s3a.fast.upload.active.blocks` does not consume much memory. But it may result in a large number of blocks to compete with other filesystem operations.

We recommend a low value of `fs.s3a.fast.upload.active.blocks` — enough to start background upload without overloading other parts of the system. Then experiment to see if higher values deliver more throughput — especially from VMs running on EC2.

Optimizing S3A read performance for different file types

The S3A filesystem client supports the notion of input policies, similar to that of the POSIX `fadvise()` API call. This tunes the behavior of the S3A client to optimize HTTP GET requests for reading different filetypes. To optimize HTTP GET requests, you can take advantage of the S3A input policy option `fs.s3a.experimental.input.fadvise`.

Policy	Description
"normal"	This starts off as "sequential": it asks for the whole file. As soon as the application tries to seek backwards in the file it switches into "random" IO mode. This is not quite as efficient for Random IO as the "random" mode, because that first read may have to be aborted. However, because it is adaptive, it is the best choice if you do not know the data formats which will be read.
"sequential" (default)	Read through the file, possibly with some short forward seeks. The whole document is requested in a single HTTP request; forward seeks within the readahead range are supported by skipping over the intermediate data. This leads to maximum read throughput, but with very expensive backward seeks.
"random"	Optimized for random IO, specifically the Hadoop <code>PositionedReadable</code> operations — though <code>seek(offset); read(byte_buffer)</code> also benefits. Rather than ask for the whole file, the range of the HTTP request is set to that of the length of data desired in the <code>read</code> operation - rounded up to the readahead value set in <code>setReadahead()</code> if necessary. By reducing the cost of closing existing HTTP requests, this is highly efficient for file IO accessing a binary file through a series of <code>PositionedReadable.read()</code> and <code>PositionedReadable.readFully()</code> calls. Sequential reading of a file is expensive, as now many HTTP requests must be made to read through the file.

For operations simply reading through a file (copying, DistCp, reading gzip or other compressed formats, parsing .csv files, and so on) the sequential policy is appropriate. This is the default, so you do not need to configure it.

For the specific case of high-performance random access IO (for example, accessing ORC files), you may consider using the random policy in the following circumstances:

- Data is read using the `PositionedReadable` API.
- There are long distance (many MB) forward seeks.
- Backward seeks are as likely as forward seeks.
- There is little or no use of single character `read()` calls or small `read(buffer)` calls.
- Applications are running close to the Amazon S3 data store; that is, the EC2 VMs on which the applications run are in the same region as the Amazon S3 bucket.

You must set the desired fadvise policy in the configuration option `fs.s3a.experimental.input.fadvise` when the filesystem instance is created. It can only be set on a per-filesystem basis, not on a per-file-read basis. You can set it in `core-site.xml`:

```
<property>
  <name>fs.s3a.experimental.input.fadvise</name>
  <value>random</value>
</property>
```

Or, you can set it in the `spark-defaults.conf` configuration of Spark:

```
spark.hadoop.fs.s3a.experimental.input.fadvise random
```

Be aware that this random access performance comes at the expense of sequential IO — which includes reading files compressed with gzip.

S3 Performance Checklist

There are various types of checklists that you can use to ensure optimal performance when working with data in S3.

Checklist for Data

- [] Amazon S3 bucket is in same region as the EC2-hosted cluster.
- [] The directory layout is "shallow". For directory listing performance, the directory layout prefers "shallow" directory trees with many files over deep directory trees with only a few files per directory.
- [] The "pseudo" block size set in `fs.s3a.block.size` is appropriate for the work to be performed on the data.
- [] Copy to HDFS any data that needs to be repeatedly read to HDFS.

Checklist for Cluster Configurations

- [] Set `yarn.scheduler.capacity.node-locality-delay` to 0 to improve container launch times.

Checklist for Code

- [] Application does not make `rename()` calls. Where it does, it does not assume the operation is immediate.
- [] Application does not assume that `delete()` is near-instantaneous.
- [] Application uses `FileSystem.listFiles(path, recursive=true)` to list a directory tree.
- [] Application prefers forward seeks through files, rather than full random IO.
- [] If making "random" IO through `seek()` and `read()` sequences or and Hadoop's `PositionedReadable` API, `fs.s3a.experimental.input.fadvise` is set to `random`.

Troubleshooting S3 and S3Guard

It can be a bit troublesome to get the S3A connector to work, with classpath and authentication being the usual trouble spots. Use the tips provided in the following links to troubleshoot errors.

S3: <https://aws.amazon.com/premiumsupport/knowledge-center/>

S3Guard: https://hadoop.apache.org/docs/current3/hadoop-aws/tools/hadoop-aws/troubleshooting_s3a.html

Related Information

[Encrypting Data on S3](#)