AmazonS3

S3 Support in Apache Hadoop

Apache Hadoop ships with a connector to S3 called "S3A", with the url prefix "s3a:"; its previous connectors "s3", and "s3n" are deprecated and/or deleted from recent Hadoop versions.

1. Consult the Latest Hadoop documentation for the specifics on using any the S3A connector.
2. For Hadoop 2.x releases, the latest troubleshooting documentation.
3. For Hadoop 3.x releases, the latest troubleshooting documentation.

S3 Support in Amazon EMR

Amazon's EMR Service is based upon Apache Hadoop, but contains modifications and their own closed-source S3 client. Consult Amazon's documentation on this. Only Amazon can provide support and/or field bug reports related to their S3 support.

Important: Classpath setup

1. The S3A connector is implemented in the hadoop-aws JAR. If it is not on the classpath: stack trace.
2. Do not attempt to mix a "hadoop-aws" version with other hadoop artifacts from different versions. They must be from exactly the same release. Otherwise: stack trace.
3. The S3A connector is depends on AWS SDK JARs. If they are not on the classpath: stack trace.
4. Do not attempt to use an amazon S3 SDK JAR different from the one which the hadoop version was built with. Otherwise: stack trace highly likely.
5. The normative list of dependencies of a specific version of the hadoop-aws JAR are stored in Maven, which can be viewed on mvnrepsitory.

Important: you need a consistency layer to use Amazon S3 as a destination of MapReduce, Spark and Hive work

You cannot use any of the S3 filesystem clients as a drop-in replacement for HDFS. Amazon S3 is an "object store" with

- Eventual consistency: changes made by one application (creation, updates and deletions) will not be visible until some undefined time.
- Non-atomic rename and delete operations. Renaming or deleting large directories takes time proportional to the number of entries -and visible to other processes during this time, and indeed, until the eventual consistency has been resolved. This breaks the commit protocol used by all these applications to safely commit the output of multiple tasks within a job.

Hadoop 3.x ships with S3Guard for consistency, and the S3A Committers for committing work.

For Amazon EMR, use "Consistent EMR" for a consistent view of the store.

Note: third party S3-compatible stores may not have this limitation. Consult their documentation.

Important: Security

Your Amazon Secret Access Key is that: secret. If it gets known you have to go to the Security Credentials page and revoke it. Try and avoid printing it in logs, or checking the XML configuration files into revision control.

1. Do not ever check it in to revision control systems.
2. Although the clients (currently) support embedding the credentials in the URI, this is very dangerous: it will appear in logs and error messages. Avoid this.
3. S3A will automatically get credentials of the current IAM Role when running on an EC2 VM.