Archive Splunk indexes to Hadoop in Splunk Web
Archive Splunk indexes to Hadoop in Splunk Web

Before you begin, note the following:

- You must configure a Hadoop provider.
- Splunk must be installed using the same user for all indexers and Splunk Enterprise instances. This is the user which connects to HDFS for archiving and the user and user permissions must be consistent.
- The data in the referring Index must be in warm, cold, or frozen buckets only.
- The Hadoop client libraries must be in the same location on each indexer. Likewise, the Java Runtime Environment must be installed in the same location on each indexer. See System and software requirements for updated information about the required versions.
- The Splunk user associated with the Splunk indexer must have permission to write to the HDFS node.
- Splunk cannot currently archive buckets with raw data larger than 5GB to S3. You can configure your Splunk Enterprise bucket sizes in indexes.conf. See Archiving Splunk indexes to S3 in this manual for known issues when archiving to S3.

Configure index archiving with the user interface

1. Navigate to Settings > Virtual Indexes and select the Archived Indexes tab. You can edit any existing archived index by clicking the arrow to its left.

2. Click New Archived Indexes to archive another index.

3. Type the names of the indexes you want to archive. You can add multiple indexes. Indexes that are already archived are disabled in the drop down list.

4. Provide a suffix for the new archive indexes. For example, if you select the "_archive" suffix, the new archived index will be "indexname_archive".

5. Select the Hadoop Provider that the new archived indexes will be assigned to. Note you can determine the bandwidth by provider that these archives can use. See "Set bandwidth limits for archiving" in this topic.

6. For Destination path in HDFS, provide the path to the working directory your provider should use for this data. For example: /user/root/archive/splunk_index_archive. If you are copying data to S3, prefix
7. Determine the age of the data that is copied to the archived index. For example, if you select "5 Days," data is copied from the warm, cold, or frozen bucket in the indexer to the archive bucket when it is five days old. **Note:** Splunk deletes data after a period of time defined in your indexer settings, so make sure that this field is set to copy the buckets before they are deleted.

**Set bandwidth limits for archiving**

If you have concerns about the bandwidth required for consistent archiving, you can set bandwidth throttling. When you set throttling for a provider, the limit you set for your provider is then applied across all indexes assigned to that provider.

**Note:** We currently cannot guarantee bandwidth limits for bucket archival to S3 file systems.

To set bandwidth throttling:

1. In the Archived Indexes tab, click on **Max bandwidth (Provider)** for the index you want to edit. This opens the Edit Provider page for that index.

2. Under "Archive Settings" check **Enable Archive Bandwidth Throttling**.

3. Enter the maximum bandwidth you want to allow for all archived indexes associated with the provider.

4. Click **Save**.