LanguageManual Archiving

Archiving for File Count Reduction

Note: Archiving should be considered an advanced command due to the caveats involved.

Overview

Due to the design of HDFS, the number of files in the filesystem directly affects the memory consumption in the namenode. While normally not a problem for small clusters, memory usage may hit the limits of accessible memory on a single machine when there are >50-100 million files. In such situations, it is advantageous to have as few files as possible.

The use of Hadoop Archives is one approach to reducing the number of files in partitions. Hive has built-in support to convert files in existing partitions to a Hadoop Archive (HAR) so that a partition that may once have consisted of 100's of files can occupy just ~3 files (depending on settings). However, the trade-off is that queries may be slower due to the additional overhead in reading from the HAR.

Note that archiving does NOT compress the files – HAR is analogous to the Unix tar command.

Settings

There are 3 settings that should be configured before archiving is used. (Example values are shown.)

```
hive> set hive.archive.enabled=true;
hive> set hive.archive.har.parentdir.settable=true;
hive> set har.partfile.size=1099511627776;
```

`hive.archive.enabled` controls whether archiving operations are enabled.

`hive.archive.har.parentdir.settable` informs Hive whether the parent directory can be set while creating the archive. In recent versions of Hadoop the `-p` option can specify the root directory of the archive. For example, if `/dir1/dir2/file` is archived with `/dir1` as the parent directory, then the resulting archive file will contain the directory structure `dir2/file`. In older versions of Hadoop (prior to 2011), this option was not available and therefore Hive must be configured to accommodate this limitation.

`har.partfile.size` controls the size of the files that make up the archive. The archive will contain *`size_of_partition`*`/har.partfile.size` files, rounded up. Higher values mean fewer files, but will result in longer archiving times due to the reduced number of mappers.

Usage

Archive

Once the configuration values are set, a partition can be archived with the command:

```
ALTER TABLE table_name ARCHIVE PARTITION (partition_col = partition_col_value, partition_col =
partiton_col_value, ...)
```

For example:

```
ALTER TABLE srcpart ARCHIVE PARTITION(ds='2008-04-08', hr='12')
```

Once the command is issued, a mapreduce job will perform the archiving. Unlike Hive queries, there is no output on the CLI to indicate process.

Unarchive

The partition can be reverted back to its original files with the unarchive command:

```
ALTER TABLE srcpart UNARCHIVE PARTITION(ds='2008-04-08', hr='12')
```

Cautions and Limitations

- In some older versions of Hadoop, HAR had a few bugs that could cause data loss or other errors. Be sure that these patches are integrated into your version of Hadoop:

  https://issues.apache.org/jira/browse/HADOOP-6591 (fixed in Hadoop 0.21.0)

  https://issues.apache.org/jira/browse/MAPREDUCE-1548 (fixed in Hadoop 0.22.0)

  https://issues.apache.org/jira/browse/MAPREDUCE-2143 (fixed in Hadoop 0.22.0)

  https://issues.apache.org/jira/browse/MAPREDUCE-1752 (fixed in Hadoop 0.23.0)

- The HarFileSystem class still has a bug that has yet to be fixed:

  https://issues.apache.org/jira/browse/MAPREDUCE-1877 (moved to https://issues.apache.org/jira/browse/HADOOP-10906 in 2014)

Hive comes with the HiveHarFileSystem class that addresses some of these issues, and is by default the value for `fs.har.impl`. Keep this in mind if you're rolling your own version of HarFileSystem:

- The default HiveHarFileSystem.getFileBlockLocations() has **no locality**. That means it may introduce higher network loads or reduced performance.
- Archived partitions cannot be overwritten with INSERT OVERWRITE. The partition must be unarchived first.
- If two processes attempt to archive the same partition at the same time, bad things could happen. (Need to implement concurrency support.)

Under the Hood

Internally, when a partition is archived, a HAR is created using the files from the partition's original location (such as `/warehouse/table/ds=1`). The parent directory of the partition is specified to be the same as the original location and the resulting archive is named 'data.har'. The archive is moved under the original directory (such as `/warehouse/table/ds=1/data.har`), and the partition's location is changed to point to the archive.