USING HADOOP WITH CEPHFS

HADOOP CONFIGURATION

This section describes the Hadoop configuration options used to control Ceph. These options are intended to be set in the Hadoop configuration file *conf/core-site.xml*.

Property	Value	Notes
fs.default.name	Ceph URI	ceph:///
ceph.conf.file	Local path to ceph.conf	/etc/ceph/ceph.conf
ceph.conf.options	Comma separated list of Ceph configuration key/value pairs	opt1=val1,opt2=val2
ceph.root.dir	Mount root directory	Default value: /
ceph.object.size	Default file object size in bytes	Default value (64MB): 67108864
ceph.data.pools	List of Ceph data pools for storing file.	Default value: default Ceph pool.
ceph.localize.reads	Allow reading from file replica objects	Default value: true

SUPPORT FOR PER-FILE CUSTOM REPLICATION

Hadoop users may specify a custom replication factor (e.g. 3 copies of each block) when creating a file. However, object replication factors are controlled on a per-pool basis in Ceph, and by default a Ceph file system will contain a pre-configured pool. In order to support per-file replication Hadoop can be configured to select from alternative pools when creating new files.

Additional data pools can be specified using the `ceph.data.pools` configuration option. The value of the option is a comma separated list of pool names. The default Ceph pool will be used automatically if this configuration option is omitted or the value is empty. For example, the following configuration setting will consider the three pools listed.

```
<property>
  <name>ceph.data.pools</name>
  <value>pool1,pool2,pool5</value>
</property>
```

Hadoop will not create pools automatically. In order to create a new pool with a specific replication factor use the `ceph osd pool create` command, and then set the `size` property on the pool using the `ceph osd pool set` command. For more information on creating and configuring pools see the RADOS Pool documentation.

Once a pool has been created and configured the metadata service must be told that the new pool may be used to store file data. A pool can be made available for storing file system data using the `ceph mds add_data_pool` command.

First, create the pool. In this example we create the `hadoop1` pool with replication factor 1.

```
ceph osd pool create hadoop1 100
ceph osd pool set hadoop1 size 1
```

Next, determine the pool id. This can be done using the `ceph osd dump` command. For example, we can look for the newly created `hadoop1` pool.

```
ceph osd dump | grep hadoop1
```

The output should resemble:

```
pool 3 'hadoop1' rep size 1 min_size 1 crush_ruleset 0...
```

where 3 is the pool id. Next we will use the pool id reference to register the pool as a data pool for storing file system data.

```
ceph mds add_data_pool 3
```

The final step is to configure Hadoop to consider this data pool when selecting the target pool for new files.

```
<property>
        <name>ceph.data.pools</name>
        <value>hadoop1</value>
</property>
```

POOL SELECTION SEMANTICS

The following semantics describe the rules by which Hadoop will choose a pool given a desired replication factor and the set of pools specified using the `ceph.data.pools` configuration option.

1. When no custom pools are specified the default Ceph data pool is used.
2. A custom pool with the same replication factor as the default Ceph data pool will override the default.
3. A pool with a replication factor that matches the desired replication will be chosen if it exists.
4. Otherwise, a pool with at least the desired replication factor will be chosen, or the maximum possible.

DEBUGGING POOL SELECTION

Hadoop will produce log file entry when it cannot determine the replication factor of a pool (e.g. it is not configured as a data pool). The log message will appear as follows:

```
Error looking up replication of pool: <pool name>
```

Hadoop will also produce a log entry when it wasn’t able to select an exact match for replication. This log entry will appear as follows:

```
selectDataPool path=<path> pool:repl=<name>:<value> wanted=<value>
```