WebHCat Installation

Table of contents

1 Procedure

Note: WebHCat was originally called Templeton. For backward compatibility the name still appears in URLs, log file names, variable names, etc.

1. Ensure that the required related installations are in place, and place required files into the Hadoop distributed cache.
2. Download and unpack the HCatalog distribution.
3. Set the `TEMPLETON_HOME` environment variable to the base of the HCatalog REST server installation. This will usually be same as `HCATALOG_HOME`. This is used to find the WebHCat (Templeton) configuration.
4. Set JAVA_HOME, HADOOP_PREFIX, and HIVE_HOME environment variables.
5. Review the configuration and update or create `webhcat-site.xml` as required. Ensure that site specific component installation locations are accurate, especially the Hadoop configuration path. Configuration variables that use a filesystem path try to have reasonable defaults, but it's always safe to specify a full and complete path.
6. Verify that HCatalog is installed and that the `hcat` executable is in the PATH.
7. Build HCatalog using the command `ant jar` from the top level HCatalog directory.
8. Start the REST server with the command `sbin/webhcat_server.sh start`.
9. Check that your local install works. Assuming that the server is running on port 50111, the following command would give output similar to that shown.

```
% curl -i http://localhost:50111/templeton/v1/status
HTTP/1.1 200 OK
Content-Type: application/json
Transfer-Encoding: chunked
Server: Jetty(7.6.0.v20120127)

{"status":"ok","version":"v1"}
%
```

2 Server Commands

- **Start the server:** `sbin/webhcat_server.sh start`
- **Stop the server:** `sbin/webhcat_server.sh stop`
- **End-to-end build, run, test:** `ant e2e`

3 Requirements

- Ant, version 1.8 or higher
- Hadoop, version 2.1.0
- ZooKeeper version 3.4.5 is required if you are using the ZooKeeper storage class. (Be sure to review and update the ZooKeeper-related WebHCat configuration.)

- HCatalog, version 0.11.0. The hcat executable must be both in the PATH and properly configured in the WebHCat configuration.
- Permissions must be given to the user running the server. (See below.)
- If running a secure cluster, Kerberos keys and principals must be created. (See below.)
- Hadoop Distributed Cache. To use Hive version 0.11.0, Pig version 0.11.1, or hadoop/streaming resources, see instructions below for placing the required files in the Hadoop Distributed Cache.

4 Hadoop Distributed Cache

The server requires some files be accessible on the Hadoop distributed cache. For example, to avoid the installation of Pig and Hive everywhere on the cluster, the server gathers a version of Pig or Hive from the Hadoop distributed cache whenever those resources are invoked. After placing the following components into HDFS please update the site configuration as required for each.

- **Hive**: Download the Hive 0.11.0 tar.gz file and place it in HDFS.

```
hadoop fs -put /tmp/hive-0.11.0.tar.gz /apps/templeton/hive-0.11.0.tar.gz
```

- **Pig**: Download the Pig 0.11.1 tar.gz file and place it into HDFS. For example:

```
hadoop fs -put /tmp/pig-0.11.1.tar.gz /apps/templeton/pig-0.11.1.tar.gz
```

- **Hadoop Streaming**:

  Place `hadoop-streaming-*.jar` into HDFS. Use the following command:

```
hadoop fs -put <hadoop streaming jar> \
  <templeton.streaming.jar>/hadoop-streaming-*.jar
```

  where <*templeton.streaming.jar*> is a property value defined in webhcat-default.xml which can be overridden in the webhcat-site.xml file, and <*hadoop streaming jar*> is the Hadoop streaming jar in your Hadoop version:

  - `hadoop-2.*/share/hadoop/tools/lib/hadoop-streaming-*.jar` in the Hadoop 2.x tar
  - `hadoop-1.*/contrib/streaming/hadoop-streaming-*.jar` in the Hadoop 1.x tar

  For example,

```
hadoop fs -put hadoop-2.1.0/share/hadoop/tools/lib/hadoop-streaming-2.1.0.jar \
  /apps/templeton/hadoop-streaming.jar
```

- **Override Jars**: Place override jars required (if any) into HDFS. *Note*: Hadoop versions prior to 1.0.3 required a patch (HADOOP-7987) to properly run WebHCat. This patch is distributed with WebHCat (located at `templeton/src/hadoop_temp_fix/ugi.jar`) and should be placed into HDFS, as reflected in the current default configuration.

```
hadoop fs -put ugi.jar /apps/templeton/ugi.jar
```

The location of these files in the cache, and the location of the installations inside the archives, can be specified using the following WebHCat configuration variables. (See the Configuration documentation for more information on changing WebHCat configuration parameters.)

Name	Default	Description
templeton.pig.archive	`hdfs:///apps/templeton/pig-0.11.1.tar.gz`	The path to the Pig archive.
templeton.pig.path	`pig-0.11.1.tar.gz/pig-0.11.1/bin/pig`	The path to the Pig executable.
templeton.hive.archive	`hdfs:///apps/templeton/hive-0.11.0.tar.gz`	The path to the Hive archive.
templeton.hive.path	`hive-0.11.0.tar.gz/hive-0.11.0/bin/hive`	The path to the Hive executable.
templeton.streaming.jar	`hdfs:///apps/templeton/hadoop-streaming.jar`	The path to the Hadoop streaming jar file.
templeton.override.jars	`hdfs:///apps/templeton/ugi.jar`	Jars to add to the HADOOP_CLASSPATH for all Map Reduce jobs. These jars must exist on HDFS.

5 Permissions

Permission must given for the user running the WebHCat executable to run jobs for other users. That is, the WebHCat server will impersonate users on the Hadoop cluster.

Create (or assign) a Unix user who will run the WebHCat server. Call this USER. See the Secure Cluster section below for choosing a user on a Kerberos cluster.

Modify the Hadoop core-site.xml file and set these properties:

Variable	Value
hadoop.proxyuser.USER.groups	A comma separated list of the Unix groups whose users will be impersonated.
hadoop.proxyuser.USER.hosts	A comma separated list of the hosts that will run the hcat and JobTracker servers.

6 Secure Cluster

To run WebHCat on a secure cluster follow the Permissions instructions above but create a Kerberos principal for the WebHCat server with the name `USER/host@realm`.

Also, set the WebHCat configuration variables `templeton.kerberos.principal` and `templeton.kerberos.keytab`.