Load and Store Interfaces

Table of contents

Copyright © 2013 The Apache Software Foundation All rights reserved.

1 Set Up

The HCatLoader and HCatStorer interfaces are used with Pig scripts to read and write data in HCatalog managed tables.

2 HCatLoader

HCatLoader is used with Pig scripts to read data from HCatalog managed tables.

2.1 Usage

HCatLoader is accessed via a Pig load statement.

```
A = LOAD 'tablename' USING org.apache.hcatalog.pig.HCatLoader();
```

Assumptions

You must specify the table name in single quotes: LOAD 'tablename'. If you are using a non-default database you must specify your input as 'dbname.tablename'. If you are using Pig 0.9.2 or earlier, you must create your database and table prior to running the Pig script. Beginning with Pig 0.10 you can issue these create commands in Pig using the SQL command.

The Hive metastore lets you create tables without specifying a database; if you created tables this way, then the database name is 'default' and is not required when specifying the table for HCatLoader.

If the table is partitioned, you can indicate which partitions to scan by immediately following the load statement with a partition filter statement (see **Examples**).

2.2 HCatalog Data Types

Restrictions apply to the types of columns HCatLoader can read.

HCatLoader can read **only** the data types listed in the table. The table shows how Pig will interpret the HCatalog data type.

HCatalog Data Type	**Pig Data Type**
primitives (int, long, float, double, string)	int, long, float, double, string to chararray
map (key type should be string, valuetype must be string)	map
List<any type>	bag
struct<any type fields>	tuple

Copyright © 2013 The Apache Software Foundation All rights reserved.

2.3 Running Pig with HCatalog

Pig does not automatically pick up HCatalog jars. You will need tell Pig where your HCatalog jars are. These include the Hive jars used by the HCatalog client. To do this, you must define the environment variable PIG_CLASSPATH with the appropriate jars. HCat can tell you the jars it needs. In order to do this it needs to know where Hadoop is installed. Also, you need to tell Pig the URI for your metastore, in the PIG_OPTS variable. In the case where you have installed Hadoop and HCatalog via tar, you can do:

```
export HADOOP_HOME=<path_to_hadoop_install>
export HCAT_HOME=<path_to_hcat_install>
export PIG_CLASSPATH=$HCAT_HOME/share/hcatalog/hcatalog-0.4.0.jar:$HIVE_HOME/lib/hive-
metastore-0.9.0.jar:
$HIVE_HOME/lib/libthrift-0.7.0.jar:$HIVE_HOME/lib/hive-exec-0.9.0.jar:$HIVE_HOME/lib/
libfb303-0.7.0.jar:
$HIVE_HOME/lib/jdo2-api-2.3-ec.jar:$HIVE_HOME/conf:$HADOOP_HOME/conf:$HIVE_HOME/lib/slf4j-
api-1.6.1.jar

export PIG_OPTS=-Dhive.metastore.uris=thrift://<hostname>:<port>

<path_to_pig_install>/bin/pig -Dpig.additional.jars=$HCAT_HOME/share/hcatalog/
hcatalog-0.4.0.jar:
$HIVE_HOME/lib/hive-metastore-0.9.0.jar:$HIVE_HOME/lib/libthrift-0.7.0.jar:$HIVE_HOME/lib/
hive-exec-0.9.0.jar:
$HIVE_HOME/lib/libfb303-0.7.0.jar:$HIVE_HOME/lib/jdo2-api-2.3-ec.jar:$HIVE_HOME/lib/slf4j-
api-1.6.1.jar <script.pig>
```

Authentication

If you are using a secure cluster and a failure results in a message like "2010-11-03 16:17:28,225 WARN hive.metastore ... - Unable to connect metastore with URI thrift://..." in /tmp/<username>/hive.log, then make sure you have run "kinit <username>@FOO.COM" to get a Kerberos ticket and to be able to authenticate to the HCatalog server.

2.4 Load Examples

This load statement will load all partitions of the specified table.

```
/* myscript.pig */
A = LOAD 'tablename' USING org.apache.hcatalog.pig.HCatLoader();
...
...
```

If only some partitions of the specified table are needed, include a partition filter statement **immediately** following the load statement in the data flow. (In the script, however, a filter statement might not immediately follow its load statement.) The filter statement can include conditions on partition as well as non-partition columns.

Copyright © 2013 The Apache Software Foundation All rights reserved.

```
/* myscript.pig */
A = LOAD 'tablename' USING  org.apache.hcatalog.pig.HCatLoader();

-- date is a partition column; age is not
B = filter A by date == '20100819' and age < 30;

-- both date and country are partition columns
C = filter A by date == '20100819' and country == 'US';
...
...
```

To scan a whole table, for example:

```
a = load 'student_data' using org.apache.hcatalog.pig.HCatLoader();
b = foreach a generate name, age;
```

Notice that the schema is automatically provided to Pig, there's no need to declare name and age as fields, as if you were loading from a file.

To scan a single partition of the table web_logs, for example, partitioned by the column datestamp:

```
a = load 'web_logs' using org.apache.hcatalog.pig.HCatLoader();
b = filter a by datestamp == '20110924';
```

Pig will push the datestamp filter shown here to HCatalog, so that HCatalog knows to just scan the partition where datestamp = '20110924'. You can combine this filter with others via 'and':

```
a = load 'web_logs' using org.apache.hcatalog.pig.HCatLoader();
b = filter a by datestamp == '20110924' and user is not null;
```

Pig will split the above filter, pushing the datestamp portion to HCatalog and retaining the `user is not null` part to apply itself. You can also give a more complex filter to retrieve a set of partitions.

Filter Operators

A filter can contain the operators 'and', 'or', '()', '==', '!=', '<', '>', '<=' and '>='.

For example:

```
a = load 'web_logs' using org.apache.hcatalog.pig.HCatLoader();
b = filter a by datestamp > '20110924';
```

A complex filter can have various combinations of operators, such as:

```
a = load 'web_logs' using org.apache.hcatalog.pig.HCatLoader();
```

Copyright © 2013 The Apache Software Foundation All rights reserved.

```
b = filter a by datestamp == '20110924' or datestamp == '20110925';
```

These two examples have the same effect:

```
a = load 'web_logs' using org.apache.hcatalog.pig.HCatLoader();
b = filter a by datestamp >= '20110924' and datestamp <= '20110925';


a = load 'web_logs' using org.apache.hcatalog.pig.HCatLoader();
b = filter a by datestamp <= '20110925' and datestamp >= '20110924';
```

3 HCatStorer

HCatStorer is used with Pig scripts to write data to HCatalog managed tables.

3.1 Usage

HCatStorer is accessed via a Pig store statement.

```
A = LOAD ...
B = FOREACH A ...
...
...
my_processed_data = ...

STORE my_processed_data INTO 'tablename' USING
 org.apache.hcatalog.pig.HCatStorer();
```

Assumptions

You must specify the table name in single quotes: LOAD 'tablename'. Both the database and table must be created prior to running your Pig script. If you are using a non-default database you must specify your input as 'dbname.tablename'. If you are using Pig 0.9.2 or earlier, you must create your database and table prior to running the Pig script. Beginning with Pig 0.10 you can issue these create commands in Pig using the SQL command.

The Hive metastore lets you create tables without specifying a database; if you created tables this way, then the database name is 'default' and you do not need to specify the database name in the store statement.

For the USING clause, you can have a string argument that represents key/value pairs for partition. This is a mandatory argument when you are writing to a partitioned table and the partition column is not in the output column. The values for partition keys should NOT be quoted.

If partition columns are present in data they need not be specified as a STORE argument. Instead HCatalog will use these values to place records in the appropriate partition(s). It is

Copyright © 2013 The Apache Software Foundation All rights reserved.

valid to specify some partition keys in the STORE statement and have other partition keys in the data.

3.2 Store Examples

You can write to non-partitioned table simply by using HCatStorer. The contents of the table will be overwritten:

```
store z into 'student_data' using org.apache.hcatalog.pig.HCatStorer();
```

To add one new partition to a partitioned table, specify the partition value in store function. Pay careful attention to the quoting, as the whole string must be single quoted and separated with an equals sign:

```
store z into 'web_data' using org.apache.hcatalog.pig.HCatStorer('datestamp=20110924');
```

To write into multiple partitions at one, make sure that the partition column is present in your data, then call HCatStorer with no argument:

```
store z into 'web_data' using org.apache.hcatalog.pig.HCatStorer();
  -- datestamp must be a field in the relation z
```

3.3 HCatalog Data Types

Restrictions apply to the types of columns HCatStorer can write.

HCatStorer can write **only** the data types listed in the table. The table shows how Pig will interpret the HCatalog data type.

HCatalog Data Type	**Pig Data Type**
primitives (int, long, float, double, string)	int, long, float, double, string to chararray
map (key type should be string, valuetype must be string)	map
List<any type>	bag
struct<any type fields>	tuple

Copyright © 2013 The Apache Software Foundation All rights reserved.