Hadoop Map-Reduce Connector

Plugin module	Data provider name	Automatic metadata detection
HADOOP	HADOOPMAP	No

The Hadoop Map-Reduce Connector allows external tables to fetch data from a Hadoop filesystem by submitting a map-reduce job to convert and filter the data records and return them to the server.

Prerequisites

- The Hadoop client software and a working Java Runtime Environment installed on all DB nodes
- Access to a Hadoop cluster, and the ability to submit map-reduce jobs to it
- Ability to make network connections from every DB node to every Hadoop node, and vice versa

Examples

Load the plugin

```
create module hadoop;
alter module hadoop set mode active;
```

Create a connector

```
create connector mymapcon
source hadoopmap
target 'namenode namenodeaddress:port,
jobtracker jobtrackeraddress:port';
```

Create an external table

```
create external table customer2013 (
        id int,
        name varchar(100),
        company varchar(100),
        address varchar(400),
        countrycode char(3)
)
from mymapcon
target 'file /user/customers/2013/*.csv';
```

Target string attributes

Attribute	Type	Default	Description
namenode	string	Hadoop default	The IP address and port number of the namenode of the Hadoop cluster to use, e.g. 172.30.21.1:9000. The alternative name cldbnode may also be used.
jobtracker	string	Hadoop default	The IP address and port number of the job tracker node of your Hadoop cluster. It works the same way as the namenode attribute. The alternative name resourcemanager may also be used.
file	string	none (required)	The input file in the Hadoop filesystem, or a wildcard pattern which may match many files.
subnets	comma-separated list	list of subnets of network interfaces used by Kognitio	Specify what network interfaces to listen on to communicate with the map job. This should be a comma-separated list of subnets. A subnet is a network address in CIDR notation, e.g. 172.30.21.0/24. This attribute is useful if to communicate with the Hadoop nodes a specific network interface must be used.

Additional attributes can be used to specify how the input files are formatted; see the **Target String Format Attributes** reference sheet.

Module parameters

Set module parameters using, e.g.: `alter module hadoop set parameter java_home to '/usr/share/java/jre';`

Parameter	Default	Description
hadoop_client	/usr/bin/hadoop	The path to the hadoop client application, with which the plugin will start a map-reduce job.
hadoop_home	Directory containing hadoop-core jar	The directory where most of Hadoop's files sit, so the plugin can find Hadoop's streaming jar file. If not specified, "hadoop classpath" is run and the directory is inferred from that.
hadoop_streaming	Searched for in *<hadoop home>*/contrib/streaming/ and /usr/lib/hadoop-mapreduce	Path to Hadoop streaming jar file, or directory containing file matching hadoop-streaming*.jar.
java_home	/usr/lib/jvm/jre	The path to the Java Runtime Environment.
mapreduce_command_args		Any command-line arguments to supply to wxconverter, which is the map-reduce task. If you set this to -l -v then debug output will appear in /tmp/wxconverter-date-time-pid.log on the Hadoop nodes.

Notes

On version 8.1, conversion errors only appear in the serverdbg file. On version 8.2 and later, they appear in SYS.IPE_CONV_ERROR as with other connectors.