openPDC HdfsBridge: Setup / Operation

Author: Josh Patterson (jpatterson@floe.tv)

Summary

- Setup and operation guide for HdfsBridge
- Used as the connecting process between the openPDC historian and HDFS

Introduction

The HdfsBridge acts as a gateway between the openPDC Historian and Hadoop's distributed File System (HDFS). This provides for a seamless off-loading mechanism between sensor collection and time series data archival in Hadoop. Although HdfsBridge is a completely from scratch written project, the authors used the Hdfs-Over-Ftp project as a reference design. The Hdfs-Over-Ftp project is located at:

https://sites.google.com/a/iponweb.net/hadoop/Home/hdfs-over-ftp

One of the primary differences functionally is the addition of a FTP command for calling the checksum mechanism in HDFS through HdfsBridge. The openPDC uses this functionality to check the transferred archive's hdfs-checksum, compute the checksum locally, and then compare the results. Pinal Patel of the openPDC project implemented HDFS's custom checksumming algorithm in the openPDC in order to produce a comparable hash. This was done to ensure reliable transfer of openPDC archival data.

Setup of HdfsBridge

Get HdfsBridge

Download the latest HdfsBridge project from http://openpdc.codeplex.com.

Setup

In the file conf/HdfsBridge-env.sh you should update the JAVA_HOME variable to point to the home of your java installation. Then edit HdfsBridge-site.xml and at the minimum you need to set the “hadoop.hdfs.uri” variable to the location of your Hadoop Cluster. An example value of this would be

hdfs://myhadoopcluster:9000

You can also set the main port and the data ports in this file as well. The basis of this project is the Mina FTP Project located at:

http://mina.apache.org/ftpserver/

Two of the configuration files in our conf directory are directly associated with the Mina FTP Project:

- ftp.jks – password file
- users.conf – user accounts file

The conf directory also contains a log4j.properties file which holds the settings for the logging system. You can configure the users.conf file to add new users for access to the Hadoop cluster over FTP.

Operation of HdfsBridge

Usage of HdfsBridge is virtually the same as using a normal FTP server from the client's perspective. You can connect to HDFS from any standard FTP client such as WinSCP.

HDFS Checksumming

In order to perform an hdfs-style checksum remotely over HdfsBridge the client application needs to be able to send a custom command. In WinSCP we'd bring down the console as shown in Figure 1 below.

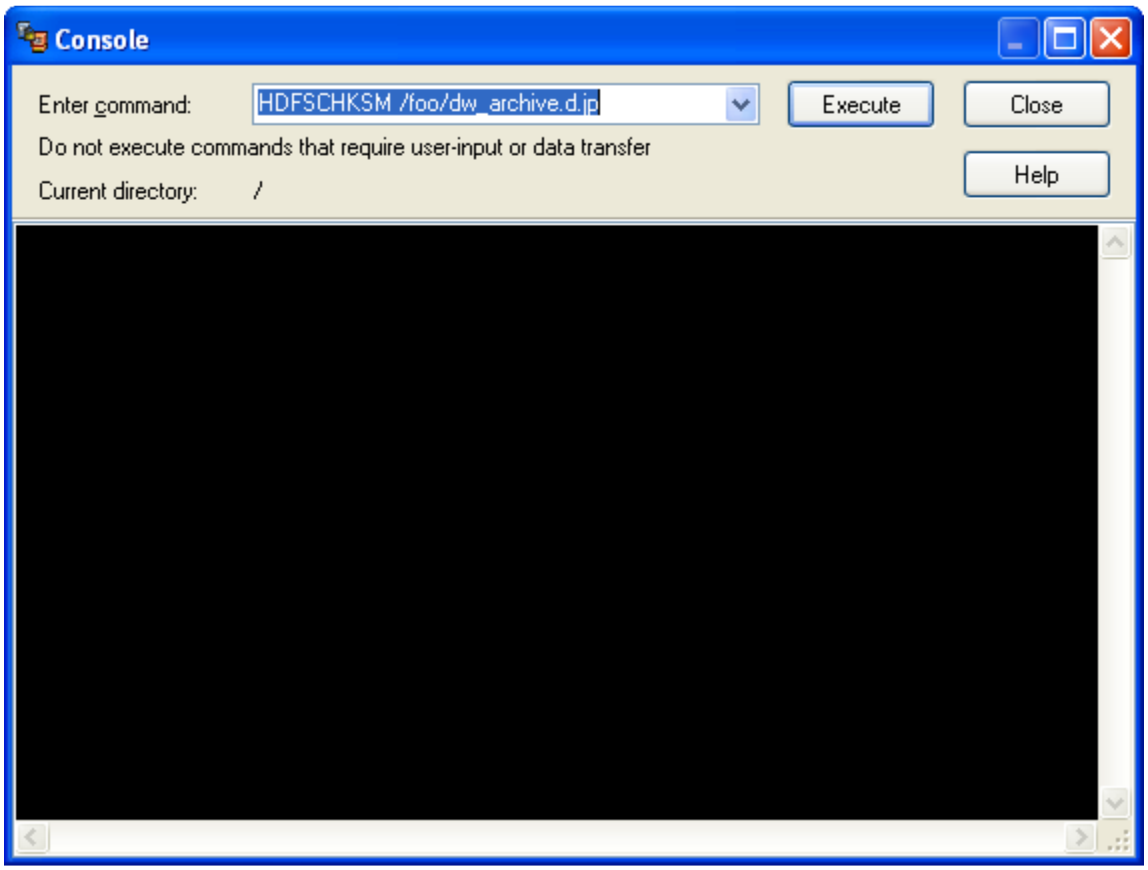

Figure 1: *The WinSCP Console Window*

In the command area we want to enter the command of the form:

HDFSCHKSM {hdfs_path}

If the checksum is completed successfully on the remote HDFS cluster a 200 code should be returned along with the resulting hash as illustrated below in Figure 2:

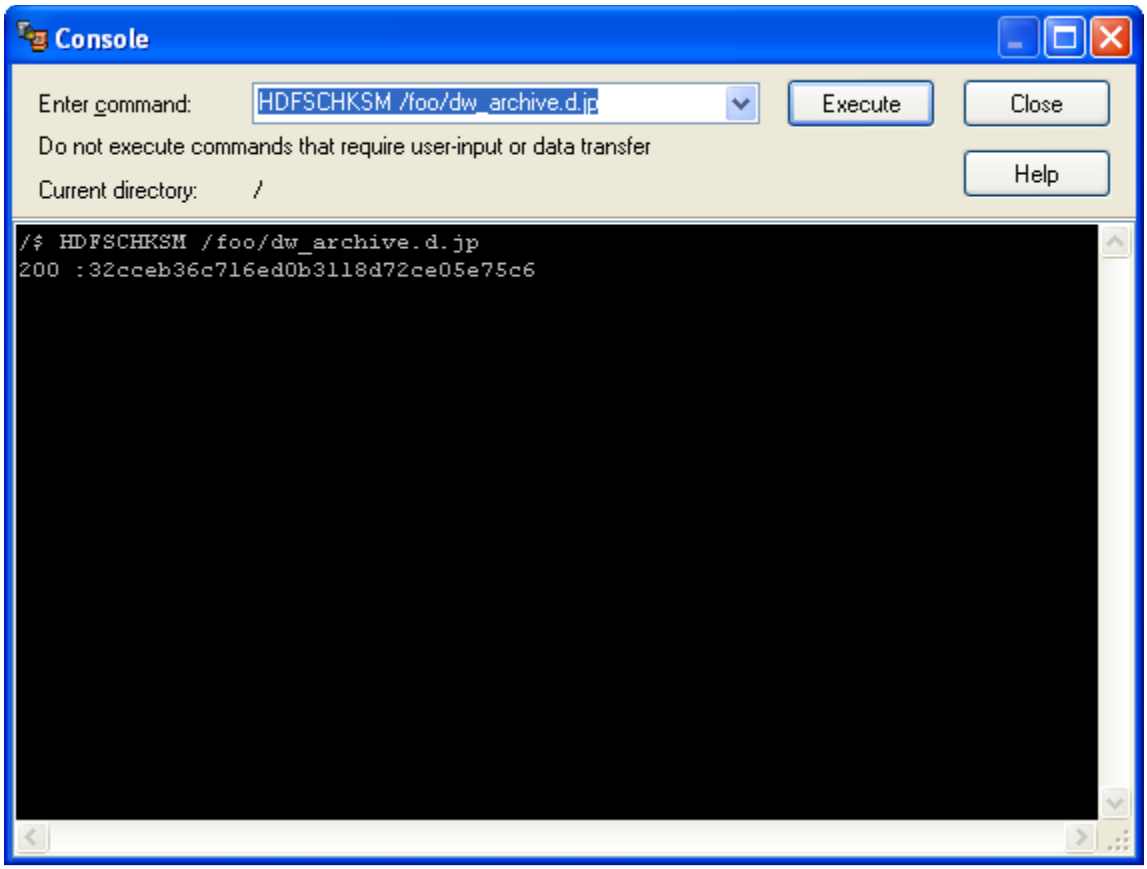

Figure 2: *Calling the HDFSCHKSM command*

Most FTP client APIs will support the ability to send a custom command