1. Introduction

Ssh-watcher is a java written application to copy files from a remote directory to a Hadoop file system.

It currently supports two protocols : FTP and SFTP

The main class to call this tool is com.ssh.hdp.Watcher which takes as an argument the path to the configuration file of the tool.

1. Configuration file

|  |  |
| --- | --- |
| Property | Definition |
| userName | username to access remote server |
| userPass | password |
| hostName | e.g: nex-sftp-01.nexius.com |
| remotePath | path to directory where files exist  e.g: /ftp\_data/sftp\_user/WRD |
| localPersistPath | Path to store the state file  = /home/albert.aouad/FTP/stateLog |
| pattern | REGEX pattern for file masks  e.g: ([^\s]+(\.(?i)(csv|stat))$) |
| localTempDir | Path on local file system to store temporary files  e.g: /home/albert.aouad/FTP/temp |
| hdfsURI | URI of Hadoop file system  e.g: hdfs://nex-hdp-14.nexius.com |
| hdfsdir | Path on HDFS where files should be loaded  e.g: /user/albert.aouad/FTP |
| logdir | Log file directory path on local filesystem  e.g: /home/albert.aouad/FTP/temp/log |
| protocol | Remote transfer protocol  Options : FTP / SSH |
| Ignore\_state | Used to ignore state file and be able to reload same file when set to true.  Options: true/false |
| hdfs.user | User which permissions will be used to write to hdfs |
| source.file.action | RENAME | DELETE | NOTHING  Action to be taken on the source file after the process finish |

1. Running the tool on Hadoop cluster

Use the following command :

hadoop jar ssh-watcher.jar com.ssh.hdp.Watcher /home/albert.aouad/FTP/ssh-watcher.properties

ssh-watcher.jar : jar package of the tool after being compile acd packaged using Eclipse (maven package)

com.ssh.hdp.Watcher : FQCN of the main class to launch the tool

/home/albert.aouad/FTP/ssh-watcher.properties : path to configuration file

1. Using REGEX Pattern

The image below explains the use of regex to match filemasks.

Machine generated alternative text:
Image File Extension Regular Expression Pattern
([“\s]+(\. (?i) (jpglpnglgiflbmp))$)
Description
#Start of the group #1
t must contains one or more anything (except white space)
( * start of the group *2
\. * follow by a dot “.,‘
(?i) * ignore the case sensive checking for the following characters
( * start of the group #3
jpg * contains characters “jpg”
* ..or
png * contains characters “png”
. . or
gif t contains characters “gif”
* ..or
bmp * contains characters “bmp”
) t end of the group *3
* end of the group #2
$ t end of the string
*end of the group #1
Whole combination is means, must have I or more strings (but not white space). follow by dot and string end in ‘jpg’ or
“png” or “gil’ or “bmp”. and the file extensive is case-insensitive