

Universal Task Documentation

Universal Automation Center support for HDFS file Monitoring and triggering

ut-hdfs-file-monitor-linux

Associated Activities:

Date: 15 October 2018

Author: Nils Buer

Revision: 00

CONFIDENTIALITY INFORMATION

Distribution list: Stonebranch Customer

Revision	Date	Author	Changes
00	20181015	Nils Buer	Initial Document (WIP)

Abstract:

This Universal Task provides a file monitor and optionally trigger, which monitors a file on the Hadoop hdfs file system. If the file is found the file monitor goes to success or can optionally launch automatically a Universal Controller Task. It also possible to run the Task monitor in auto restart mode, which means it re-starts itself automatically after it found a file.

Contents

1	Disclaimer	3
2	Scope	3
3	Introduction.....	3
4	Installation.....	3
4.1	<i>Software Requirements for Linux Agent</i>	<i>3</i>
4.2	<i>Installation Steps</i>	<i>4</i>
5	Universal Task Configuration	5
6	Test Cases	8
7	Document References.....	11

1 Disclaimer

No support and no warranty are provided by Stonebranch GmbH for this document and the related Universal Task. The use of this document and the related Universal Task is on your own risk.

Before using this task in a production system, please perform extensive testing.

Stonebranch GmbH assumes no liability for damage caused by the performance of the Universal Tasks

2 Scope

This document provides a documentation how to install and use the Universal Tasks for hdfs file monitoring and triggering.

3 Introduction

This Universal Task provides a file monitor and optionally trigger, which monitors a file on the Hadoop hdfs file system. If the file is found the file monitor goes to success or can optionally launch automatically a Universal Controller Task. It also possible to run the Task monitor in auto restart mode, which means it re-starts itself automatically after it found a file.

Some details about the universal task to monitor a value in a database column

- The Universal Task uses the Python hdfs module, which calls the Hadoop WebHDFS REST API
- The Universal Task supports both Universal Agent for Linux/Unix and Windows however it has been currently only tested against a Linux Agent
- You can select different log-levels e.g. Info and debug
- All Passwords are encrypted using Controller Credentials
- Currently only the InsecureClient (the default) is implemented the TokenClient can be implemented on request
- Currently only direct file matches are implemented as this is the standard functionality supported by the hdfs python module. A scan for files using wildcards or even regular expression can be implemented on request.

4 Installation

4.1 *Software Requirements for Linux Agent*

Universal Template name: UT-HDFS-FILE-MONTIOR-LINUX

Related UAC XML Files for template and task: [1]

Requirements :

- Python 3.6
- For Python the following modules are required:

- *os*, for linux command execution
- *datetime* and *time* to compare file timestamps
- *hdfs* to perform the hdfs file operations
- *logging*, for python loglevel support

Note: Only the module hdfs needs to be added via python installer

⇒ *pip install hdfs*

- Universal Controller V6.4.7.0 or higher
- Universal Agent V6.5.0.0 or higher installed on a Linux Server

4.2 Installation Steps

The following describes the installation steps:

1. Check the current Python Version

python -V (Note: Capital "V")

If your Version is Python 3.6 or later all is fine. If a no python or a lower Version has been installed upgrade your python Version or install the Universal Agent with the Python binding option (*--python yes*). This option will install python 3.6. along with your universal agent.

e.g.

```
sudo sh ./unvinst --network_provider oms --oms_servers 7878@192.168.88.12 --oms_port 7878
--oms_autostart no --ac_netname OPSAUTOCONF --opscli yes --python yes
```

2. Add the required python modules

In a command shell run as *sudo* or *root*:

- For Python the following modules are required:
 - *pip install hdfs*
or in case of universal Agent with python binding:
/opt/universal/python3.6/bin/python3 -m pip install hdfs
- Only run these if not available already:
- *pip install datetime*
 - *pip install logging*
 - *pip install os*

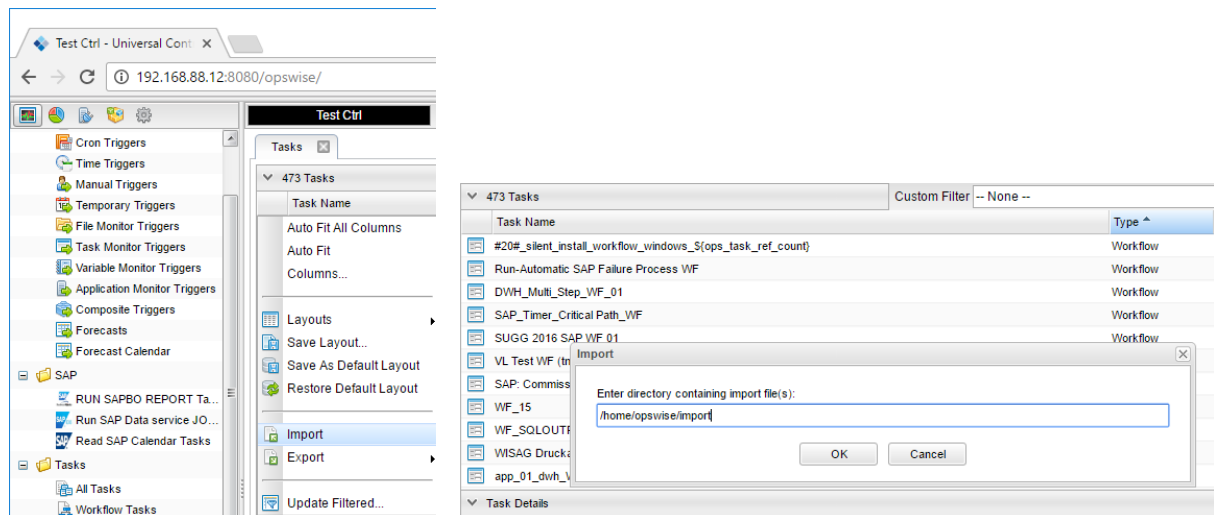
Note:

It is assumed that the modules *logging*, *sys*, *datetime*, *os* are already available. If not install them via *pip*. Only the module *hdfs* is usually not part of your installation.

<https://support.microsoft.com/en-us/help/2977003/the-latest-supported-visual-c-downloads#bookmark-vs2013>

3. Import the Universal Task including the Universal Template to your Controller

Go to "All Tasks" and load via the Import functionality the Universal Task configuration into the Controller.



5 Universal Task Configuration

1. Activate: Resolvable Credentials in Universal Automation Center properties:

Dashboards X RUN SAPBO REPORT Tasks X Properties X	
93 Properties	
Name	Value
Resolvable Credentials Permitted	true

2. Fill Out the Universal Task for hdfs file monitoring:

hdfs file monitor Task Details: Scan for file smartphone.csv

Update Launch Task View Parents Copy Delete Refresh Close

hdfs file monitor Task Variables Actions Virtual Resources Mutually Exclusive Instances Triggers Notes Versions

General

Task Name : Scan for file smartphone.csv Version : 10

Task Description :

Member of Business Services :

Resolve Name Immediately : ☐ Time Zone Preference : -- System Default --

Hold on Start : ☐

Virtual Resource Priority : 10 Hold Resources on Failure : ☐

Further Info's :

System :

hdfs file monitor Details

Agent : quickstart.cloudera - AGNT0049 Agent Cluster :

Agent Variable : ☐ Agent Cluster Variable : ☐

Credentials : Cluster Broadcast :

Credentials Variable : ☐

loglevel : DEBUG opscldir : /opt/universal/opscli/bin/

oms credentials : bigdata_cloud omshost : 192.168.88.12

omsport : 7878 hdfsuser : root

hdfsport : 50070 hdfsurl : http://localhost

file : smartphone.csv autorestart : Yes

triggeronexits : Yes tasktolaunch : Sleep30

interval : 10

Credential for Universal Task:

Credential Details: bigdata_cloud

Update Convert... Delete Refresh Close

Credential Versions

Details

Name : bigdata_cloud Version :

Type : Resolvable

Runtime User : nbuer

Runtime Password :

Description:

Field	Description
Agent	The Agent that runs the Python script assigned to the Universal Task
Loglevel	log level: DEBUG, INFO, WARNING, ERROR, CRITICAL
opsclidir	Directory where agent CLI executables are located default is: /opt/universal/opscli/bin/
omscred	<i>Credentials of the Universal Controller</i>
omshost	<i>Host of the OMS</i>
omsport	<i>Port of the OMS – default 7878</i>
hdfsuser	<i>user – User default. Defaults to the current user's (as determined by whoami). will be prefixed to all HDFS paths passed to the client</i>
hdfsport	<i>WebHDFS port on namenode default is 50070</i>
hdfsurl	<i>url – Hostname or IP address of HDFS namenode, prefixed with protocol, followed by WebHDFS port on namenode</i>
file	<i>File to monitor</i>
autorestart	<i>Defines if the monitor re-starts itself after going to success. Do not use this functionality in a Workflow Task.</i>
triggeronexits	<i>Defines if the monitor triggers also for files older than the start date of the hdfs filemonitor</i>
tasktolaunch	<i>Name of the task to launch. Note: also wildcards are supported. E.g. sleep3* starts all task starting with sleep3 e.g. slepp30, sleep3, ..</i>
interval	<i>Define in which interval the hdfs file system is scanned.</i>

6 Test Cases

The following basic test cases has been performed:

Case#	Assumed behavior	Result
1) Monitor Started 2) File Copied to hdfs Parameter <ul style="list-style-type: none"> • Trigger on exists = "No" • Autorestart = "No" • Tasktolaunch = "Sleep30" 	DEBUG - Starting new HTTP connection (1): localhost:5007 DEBUG http://localhost:50070 "GET /webhdfs/v1/?user.name=root&op=GETHOMEDIRECTORY HTTP/1.1" 200 None 2018-10-15 10:35:15,529 - DEBUG - Updated root to u'/user/root'. 2018-10-15 10:35:15,529 - DEBUG - Resolved path 'smartphone.csv' to u'/user/root/smartphone.csv'. 2018-10-15 10:35:15,533 - DEBUG - http://localhost:50070 "GET /webhdfs/v1/user/root/smartphone.csv?user.name=root&op=GETFILESTATUS HTTP/1.1" 200 None 2018-10-15 10:35:15,534 - DEBUG - modificationTime:1539624915175 2018-10-15 10:35:15,534 - DEBUG - current Time :1539624895468 2018-10-15 10:35:15,534 - DEBUG - Success, file:smartphone.csv was created after start of monitor 2018-10-15 10:35:15,974 - INFO - Task launched: Sleep30	Correct
1) File Copied to hdfs 2) Monitor Started Parameter <ul style="list-style-type: none"> • Trigger on exists = "Yes" • Autorestart = "No" • Tasktolaunch = "Sleep30" 	DEBUG - file was created before start of monitor DEBUG - Success file:smartphone.csv was created before start of monitor but triggeronexists was set INFO - Task launched: Sleep30	Correct
1) Monitor Started 2) File Copied to hdfs Parameter <ul style="list-style-type: none"> • Trigger on exists = "No" 	DEBUG - Success, file:smartphone.csv was created after start of monitor INFO - Task launched: Sleep30 INFO - Task launched: Scan for file smartphone.csv	Correct

<ul style="list-style-type: none"> • Autorestart = "Yes" • Tasktolaunch = "Sleep30" 		
1) File Copied to hdfs 2) Monitor Started 3) File Copied to hdfs Parameter <ul style="list-style-type: none"> • Trigger on exists = "Yes" • Autorestart = "Yes" • Tasktolaunch = "Sleep30" 	INFO - Task launched: Sleep30 INFO - Task Scan for file smartphone.csv re-started due to autorestart set to Yes INFO - Task launched: Sleep30 INFO - Task Scan for file smartphone.csv re-started due to autorestart set to Yes ..	Correct
1) File Copied to hdfs 2) Monitor Started Parameter <ul style="list-style-type: none"> • Trigger on exists = "Yes" • Autorestart = "Yes" • Tasktolaunch = "" 	2018-10-15 13:41:08,696 - DEBUG - Resolved path 'smartphone.csv' to '/user/root/smartphone.csv'. 2018-10-15 13:41:08,701 - DEBUG - http://localhost:50070 "GET /webhdfs/v1/user/root/smartphone.csv?user.name=root&op=GETFILESTATUS HTTP/1.1" 200 None 2018-10-15 13:41:08,702 - DEBUG - modificationTime:1539635266169 2018-10-15 13:41:08,702 - DEBUG - current Time :1539636068644 2018-10-15 13:41:08,702 - DEBUG - file was created before start of monitor 2018-10-15 13:41:08,702 - DEBUG - Success file:smartphone.csv was created before start of monitor but triggeronexists was set	Correct
Wrong opscldir	2018-10-15 13:44:24,384 - ERROR - Directory not found: /opt/universal/opscli/binx/	
Wrong omscred	success CLI requires valid user credentials. opscmd-complete	Error handling needs to be added
Wrong omshost	failed Network error: ops.agent.OPSCMDP.6E6B329247CC458B83A8E83DBD4A9FD1, ConnectNetwork, connect, 113, No route to host (192.168.88.11:7878)	correct

	<p>Network error: ops.agent.OPSCMDP.6E6B329247CC458B83A8E83DBD4A9FD1, WmsgOmsConnect, WmsgOmsConnect, 10, Failed creating OMSSConnection thread: time out</p> <p>Error sending command request. Failed creating OMSSConnection thread: time out</p> <p>2018-10-15 14:17:29,432 - ERROR - ##ERROR# Error Code: 2</p>	
Wrong omsport	<p>failed</p> <p>Network error: ops.agent.OPSCMDP.6D7475DC8FA84827BCE36B1772160525, ConnectNetwork, connect, 111, Connection refused (192.168.88.12:7879)</p> <p>Network error: ops.agent.OPSCMDP.6D7475DC8FA84827BCE36B1772160525, WmsgOmsConnect, WmsgOmsConnect, 10, Failed creating OMSSConnection thread: time out</p> <p>Error sending command request. Failed creating OMSSConnection thread: time out</p> <p>2018-10-15 14:18:36,483 - ERROR - ##ERROR# Error Code: 2</p>	correct
Wrong hdfsport	<p>Failed:</p> <p>requests.exceptions.ConnectionError: HTTPConnectionPool(host='localhost', port=50071): Max retries exceeded with url: /webhdfs/v1/?user.name=root&op=GETHOMEDIRECTORY (Caused by NewConnectionError('<urllib3.connection.HTTPConnection object at 0x7fafd3182588>: Failed to establish a new connection: [Errno 111] Connection refused',))</p>	correct
Wrong hdfsurl	<p>Failed:</p> <p>requests.exceptions.ConnectionError: HTTPConnectionPool(host='localhosts', port=50070): Max retries exceeded with url: /webhdfs/v1/?user.name=root&op=GETHOMEDIRECTORY (Caused by NewConnectionError('<urllib3.connection.HTTPConnection object at 0x7f326e5ca588>: Failed to establish a new connection: [Errno -2] Name or service not known',))</p>	correct

7 Document References

This document references the following documents:

Ref#	Description
[1] XML extract of Universal Task	<div>UAC XML extract of the Universal Template and Task</div> <ul style="list-style-type: none">ops_credentials_aef94687aa984c93aa720140bba9aebf.xmlops_task_universal_acd0f489700a48e0bf37eb04e2552378.xmlops_unv_tmplt_5e326f490075445882fd767a7faca626.xml