Files Download

Description

Files Download (uv-e-filesDownload):

This DPU downloads one or more files from the defined locations. The files to be downloaded may be located at HTTP URLs, on the local file system, at certain SFTP/FTP servers, etc.

Individual files and also whole directories may be downloaded. If a directory is provided then all files and files in subdirectories are extracted.

If an internal name (file name) is specified for the downloaded entry, this name is then used as a symbolic name to internally identify the given file further on the pipeline.
If you specify a directory as an entry then this file name is used as a prefix for the individual files within that directory.

In cases where you just need to iterate and process each downloaded file in the same way, you do not need to specify a file name.

This DPU also sets virtual path metadata for each file extracted. In case of files it is equal to the file name (local file name from the file path, e.g. example.txt from a/b/c/example.txt).
In case of directories, virtual path metadata for each extracted file is equal to the relative path of the original directory.

The URI of a file may contain macro {{execId}}, which is replaced during pipeline execution with the actual pipeline execution ID.

Configuration Parameters

Name	Description	Example
List of files and directories to download	List of files and directories to be downloaded. Each entry contains location from which the file should be obtained and optionally the internal file name.	
	URI	/tmp //Document .pdf
	Username	admin
	Password	<password>
	File name	Document. pdf
Default connection timeout (ms)		20,000
Ignore TLS/SSL errors	If checked, errors with server certificate are ignored when connecting using secure connection (SSL/TLS, URL starts with https://). Wrong host name in certificate is ignored, untrusted certificate issuers are accepted, self-signed certificates are accepted. This option causes the download to be vulnerable to man-in-the-middle attack. Use with caution, it neglects security provided by TLS/SSL connection. Connecting using this option is insecure!	false
Soft failure	In case the soft failure is checked in the configuration dialog, when there is a problem processing certain VFS entry or file, warning is shown but the execution of the DPU continues. If unchecked (default), in case of problem processing any VFS entry/file, the execution fails.	true
Skip redundant input file entries	If checked, the DPU checks whether it is not trying to process certain file URIs more times (this may happen when the DPU is configured dynamically). If yes, it just skips processing of redundant entries and logs info message.	true
Wait between calls for (ms)	Number of milliseconds the DPU should wait between the HTTP calls (0 by default, thus no delays between calls)	0

Inputs and Outputs

Name	Type	Data Unit	Description	Required
output	output	FilesDataUnit	Downloaded files	✓
config	input	RdfDataUnit	Dynamic DPU configuration, see Advanced configuration	✗

Notes

Advanced configuration

It is also possible to dynamically configure the DPU over its input `config` data unit using RDF data.

Configuration samples

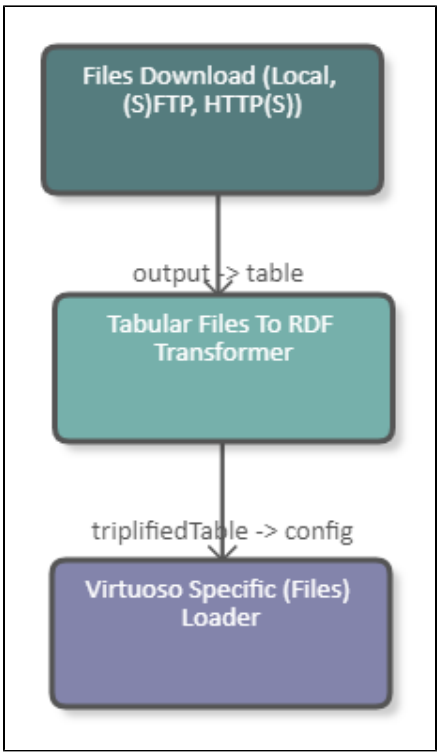
Turtle

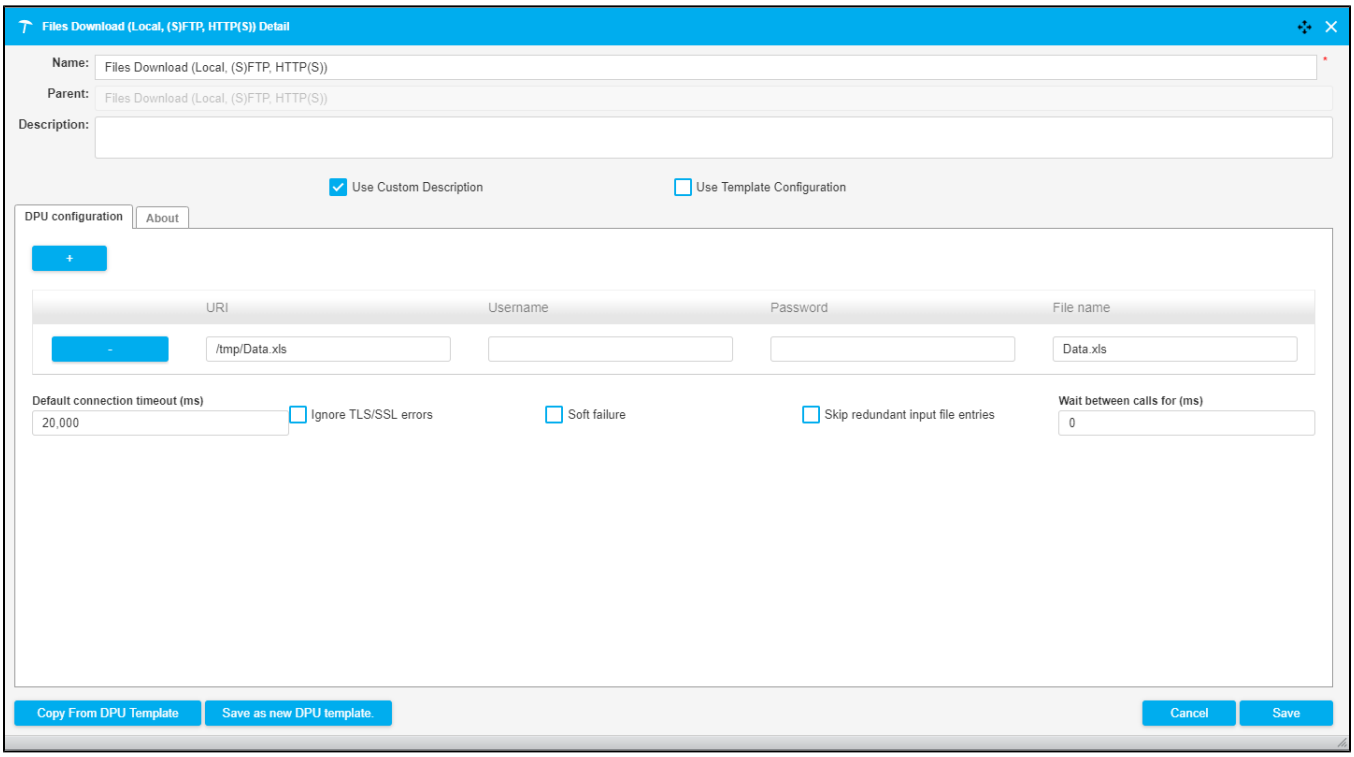
```
<http://localhost/resource/config>  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://unifiedviews.eu
/ontology/dpu/filesDownload/Config>;
        <http://unifiedviews.eu/ontology/dpu/filesDownload/hasFile> <http://localhost/resource/file/0>.


<http://localhost/resource/file/0> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://unifiedviews.eu
/ontology/dpu/filesDownload/File>;
        <http://unifiedviews.eu/ontology/dpu/filesDownload/file/uri> "http://www.zmluvy.gov.sk/data/att
/117597_dokument.pdf";
        <http://unifiedviews.eu/ontology/dpu/filesDownload/file/fileName> "zmluva.pdf".
```

Examples

Download an Excel file, convert the table data to RDF and load it to Virtuoso

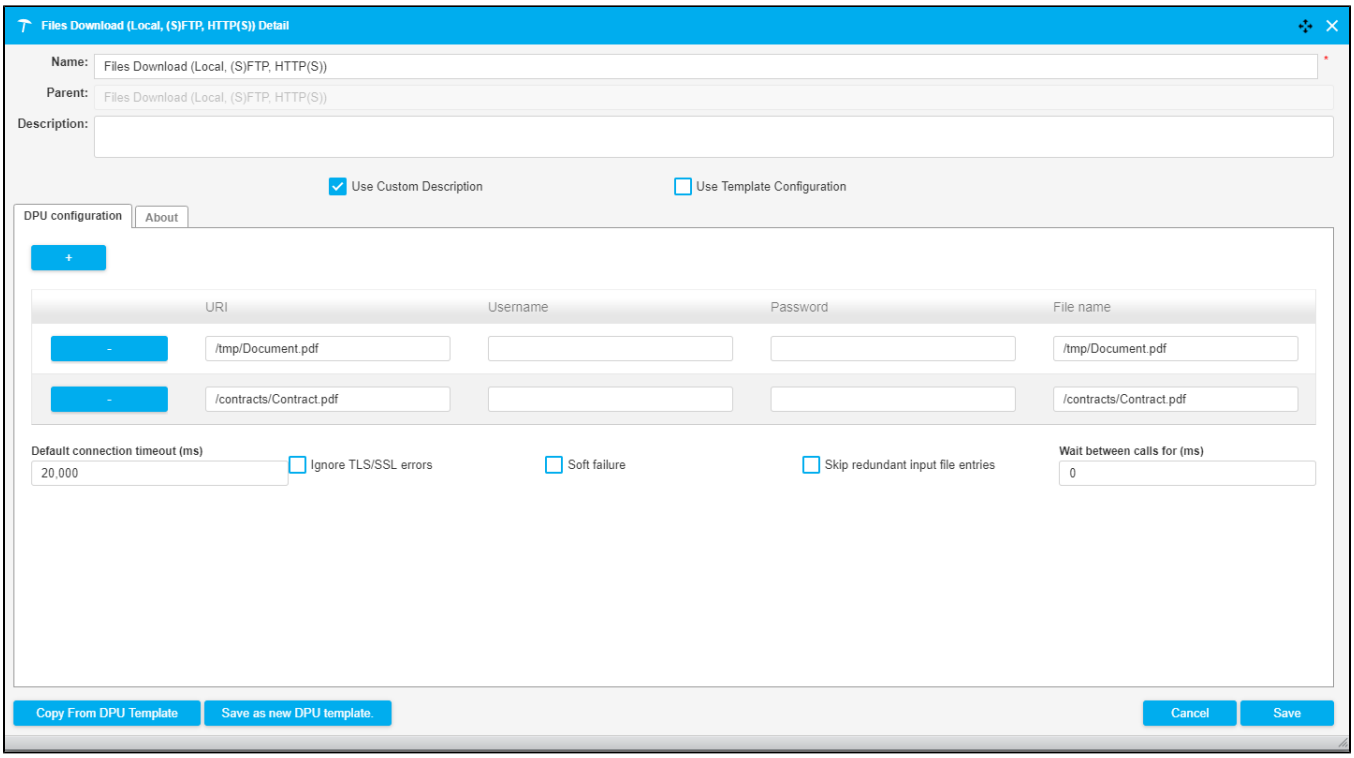
The following image shows a fragment of a pipeline which downloads an Excel file from the tmp folder of the UnifiedViews server. The data of the Excel file is subsequently converted to RDF and loaded into a Virtuoso triple store. The DPU configuration is illustrated in the image below.



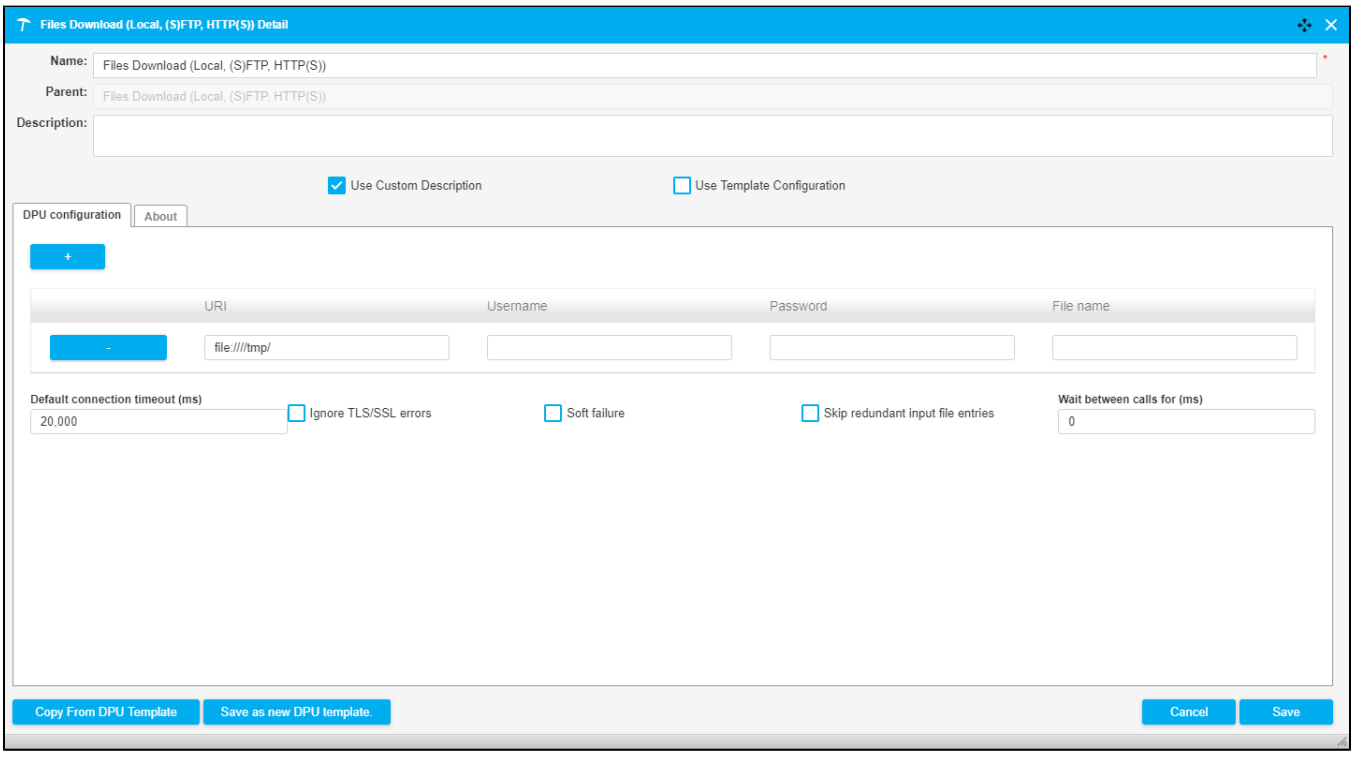


Download Multiple Files

The following image shows the configuration for downloading multiple files at once.

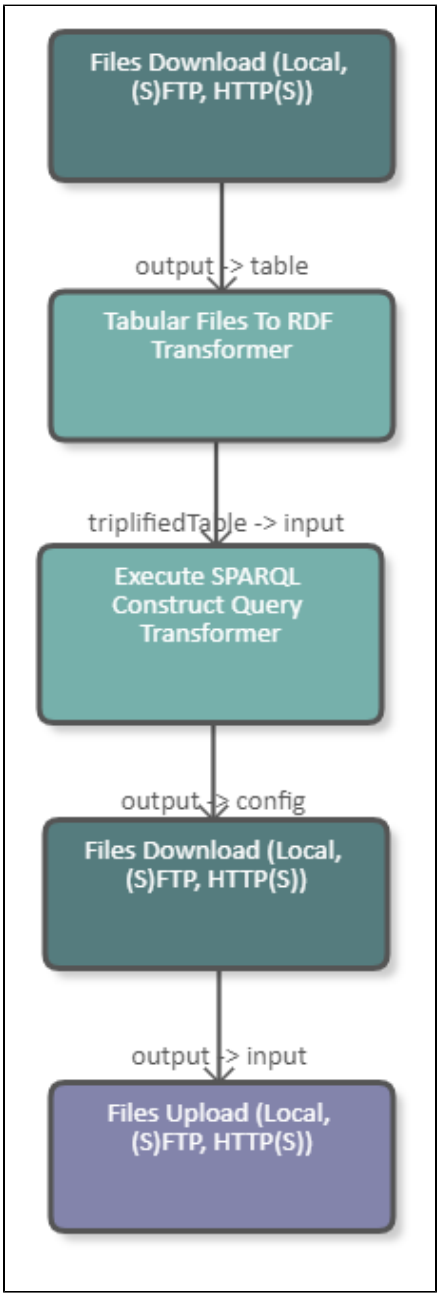


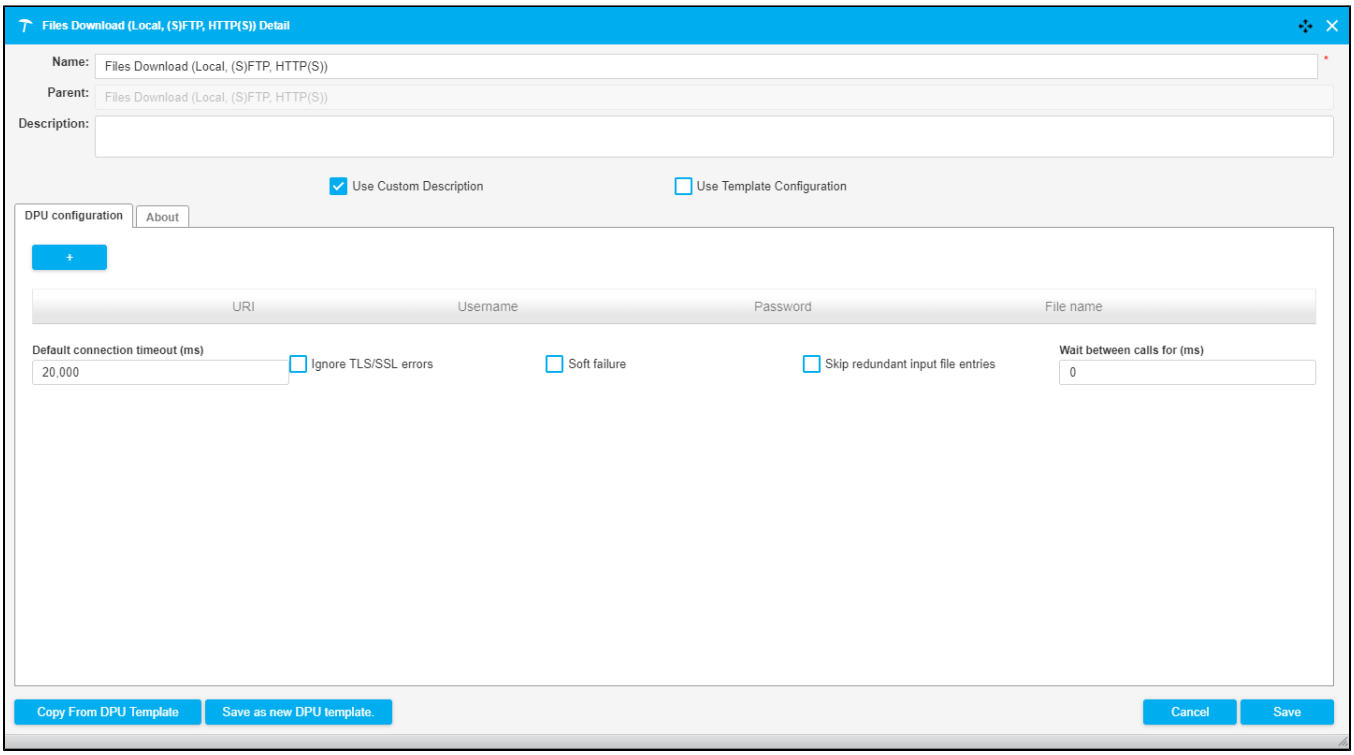
The following image shows the configuration to download all files in a directory, if any subdirectories are located here then those files will be taken as well.



Download an Excel File Containing Download Links, Convert It to RDF and Use It to Configure Another Files Download DPU

The following image shows a fragment of a pipeline which downloads an Excel file from the tmp folder of the UnifiedViews server. The data of the Excel file is subsequently converted to RDF and serves as input for a SPARQL Construct Query. The purpose of this query is to construct the configuration file of the second Files Download DPU. After the files are downloaded they are uploaded to the tmp folder of the UnifiedViews server using the Files Upload DPU. The DPU configuration is illustrated in the image below; it is empty as the configuration comes from the input RDF file.





The query used in this pipeline creates triples containing the download URI and the file name of the files that are to be downloaded. The query reads as follows:

```
CONSTRUCT {
<http://localhost/resource/config>  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://unifiedviews.eu
/ontology/dpu/filesDownload/Config>;
        <http://unifiedviews.eu/ontology/dpu/filesDownload/hasFile> <http://localhost/resource/file/0>.

<http://localhost/resource/file/0> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://unifiedviews.eu
/ontology/dpu/filesDownload/File>;
        <http://unifiedviews.eu/ontology/dpu/filesDownload/file/uri> ?fileUri;
        <http://unifiedviews.eu/ontology/dpu/filesDownload/file/fileName> ?fileName.
}
WHERE {
?s <http://localhost/fileuri/fileName> ?fileName.
?s <http://localhost/fileUri> ?fileUri
}
```