Batch downloading files with Pentaho Kettle / PDI

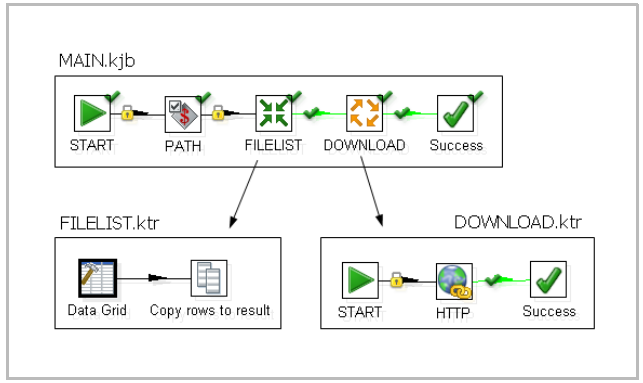
Posted on **2013/06/28**

Currently I am working on a project with the goal to download all available data sets on Eurostat as SDMX documents and load the data into a local database, so I can investigate it more conveniently. For this ETL process I use Pentaho Kettle aka PDI (Pentaho Data Integration). And one of quite a few small challenges was to download a list of files via HTTP. I must admit I was and still am surprised that this cannot be accomplished easier with Kettle – but what the heck, it's still a great tool – and first and foremost it IS possible. So in this article I am going to describe the most straightforward way to implement a batch download. I assume basic knowledge on how to use Kettle.

You can download a zip archive keeping the job here.

The MAIN job

First of all there is no transformation step to download a file via HTTP, but a job entry for that purpose, which is why we have to use a job (DOWNLOAD.kjb) within a job (MAIN.kjb). The file list will most likely have to provided by steps executed within a transformation, which extracts this information from a file. In my above mentioned use case, the list of to be downloaded files is provied within an XML document which is parsed and which offers a great opportunity for another article.



The FILELIST transformation

In this case I just use a data grid step that provides a list holding five records for five files – while one column “filename” and another column “url” specifies pretty much what you would expect. To make this table available in the next job (DOWNLOAD.kjb) I use the “Copy rows to result” step from the “Job” step-section.

The DOWNLOAD job

Because the DOWNLOAD job will download only a single file we need to launch it for every row FILELIST spits out. This you set in the job-step’s settings under “Advanced” where you have to check “Execute for every input row” – makes sense, right?

Within the job we will need the filename and the url for setting up the HTTP step. Those two fields of the incoming row we will access as variables ${URL} and ${FILENAME}. To have them available we need to take care of two things:

1) We have to declare that “URL” and “FILENAME” are parameters.

> Open the job settings by choosing Edit/Settings… while being in DOWNLOAD.kjb. In the “Paramters” section you just have to specify those two names.

2) We have to specify the field on variable/paramter mapping.

> Open the job-step settings by double-clicking on the DOWNLOAD-job-steü while being in the MAIN job. There you go again to “Paramters” and specify the stream’s column name that shall give the value for the respective variable/parameter.

Within the HTTP step we actually use the variable PATH to define the directory where the files are supposed to be saved in. This variable is specified in … yes, exactly!

The future

There will be more articles on how to get stuff done in Kettle – so you won’t have to investigate that yourself. Let’s share the wisdom! And most importantly there will be many more articles on interesting insights from Eurostat data. I can’t wait to delve right into it and dig the nuggets out. But this has to wait few more weeks as I will be on vacation in Israel until mid-July.

This entry was posted in **Pentaho** and tagged **Kettle/PDI**, **Pentaho** by **Raffael Vogler**. Bookmark the **permalink [http://www.joyofdata.de/blog/batch-downloading-files-with-pentaho-kettle/]** .

ONE THOUGHT ON “BATCH DOWNLOADING FILES WITH PENTAHO KETTLE / PDI”

Jayesh
on **2014/05/14 at 06:11** said:

This is really good explanation.. it’s works like charm.. thanks, keep posting