HULFT

Tutorial: Building a HULFT Integrate Project for Screen Scraping

Introduction to HULFT Integrate

HULFT Integrate is a flexible and powerful tool for connecting to a wide variety of data sources for reading and writing information, and provides a comprehensive set of features for manipulating, massaging, and enriching that data.

HULFT Integrate utilizes adapters to connect with data sources. Given a typical use case involving the movement and manipulation of data, we probably have the appropriate adapters available out-of-the-box to meet your requirement.

HULFT Integrate Project: Screen Scraping Use Case

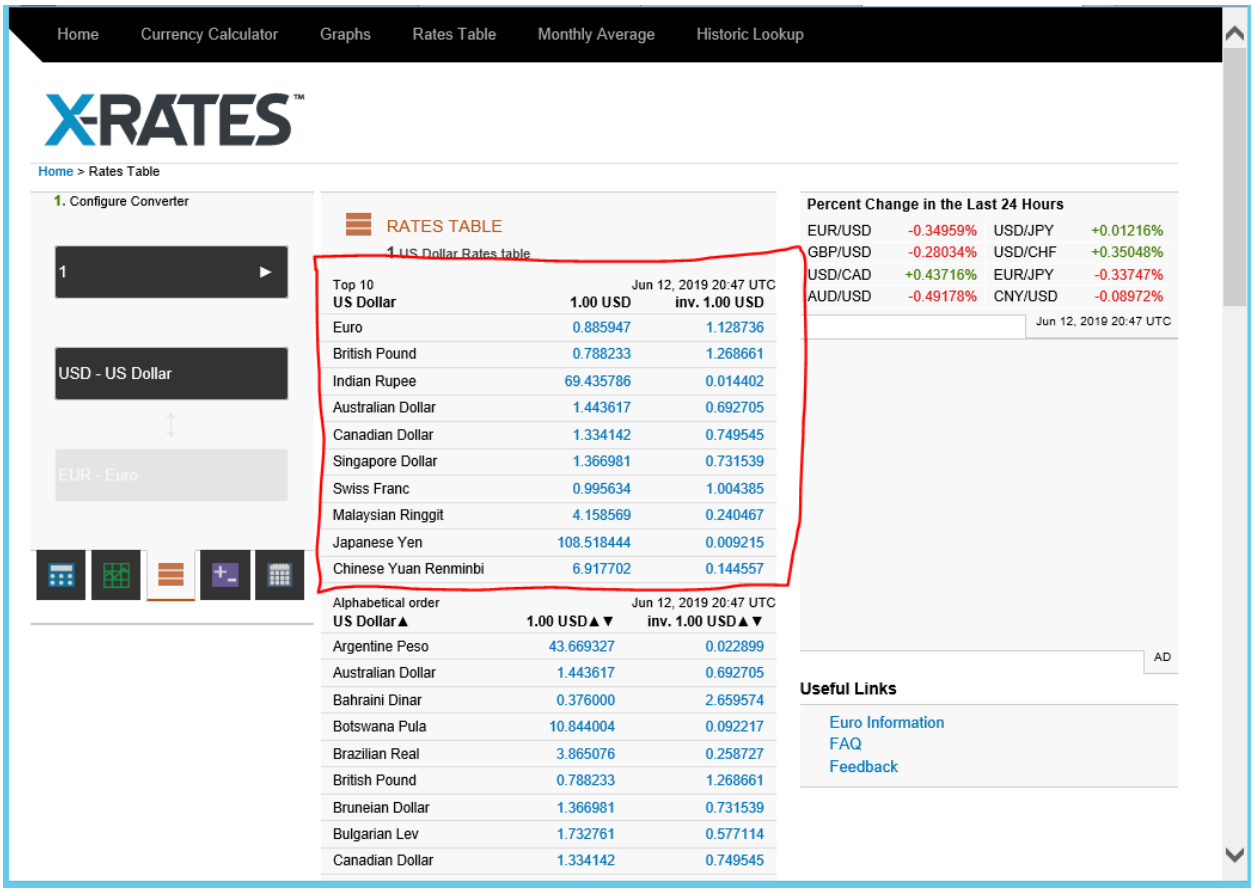
A customer had purchased a new system from a software vendor, yet still needed to extract data from the system for integration and subsequent reporting needs. The particular system in question utilized a database for a backend, but the license restricted the customer from accessing the database directly. The customer was told that an API could be built within 12 months that would allow them to extract the data they needed, but waiting up to 12 months wasn't an option. The customer needed an immediate, if not temporary, solution.

The customer decided to use a screen scraping, or web scraping approach, since the user interface was rendered in a web browser, as are many modern systems. There are many reasons why screen scraping is a poor approach from a technical perspective, and not recommended for a long-term solution. In this case, the customer was looking for a short term solution until the software vendor delivered a true API.

With this knowledge, the customer approached HULFT. Due to the flexibility of HULFT Integrate, HULFT could help with both the short term goal of screen scraping as well as the long term goal of transitioning to a true API for integration.

Building the Project Step-by-Step

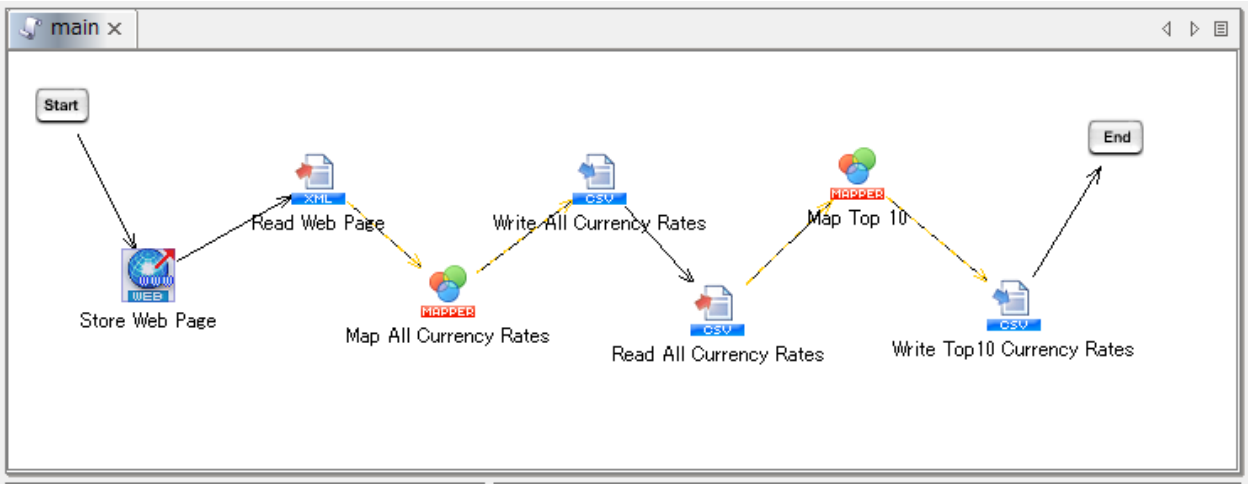
Let's use an example scenario where we want to extract the top ten currency exchange rates from a website, so we can utilize that information for further processing. Here is the website we want to extract the information from, with a highlight around the top ten exchange rates.



To create the HULFT Integrate project to accomplish this task, we need to leverage several built-in adapters from the palette;

Crawl – Allows connection to a web site and retrieval of underlying HTML
XML File Read – Allows us to read the HTML created by the Crawl adapter
CSV File Write – Allows creation of a CSV text file containing the scraped data
CSV File Read – Read a temporary CSV file created as part of the overall flow

Here is a screenshot of the completed HULFT Integrate project detailing the integration flow:



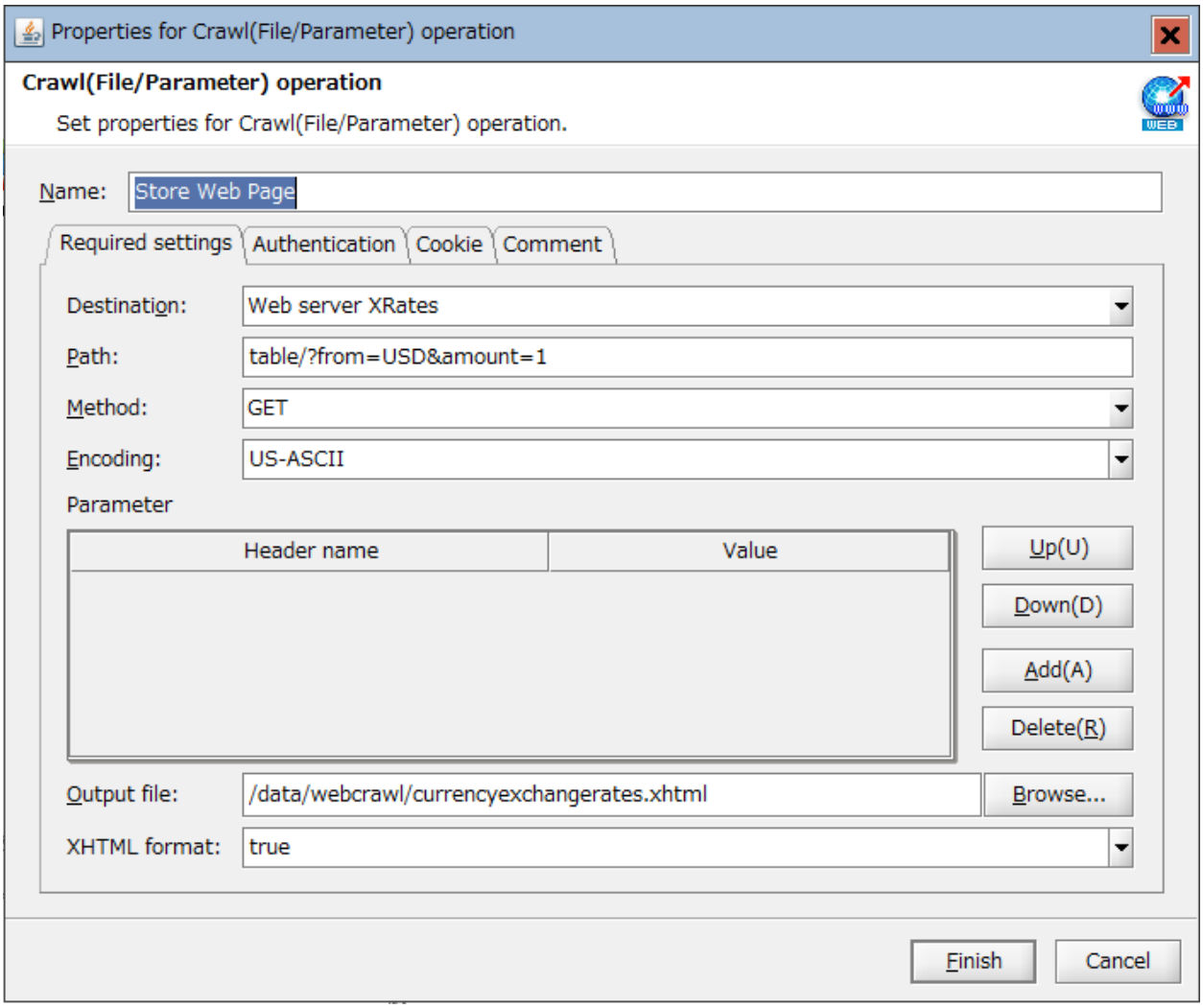
And here is a screenshot of the final CSV output showing the top ten currency exchange rates:

	A	B	C	D
1	Currency	1 USD in Cur	1 Cur in USD	
2	Euro	0.885873	1.128829	
3	British Pound	0.788187	1.268735	
4	Indian Rupee	69.39764	0.01441	
5	Australian Dollar	1.44354	0.692742	
6	Canadian Dollar	1.334085	0.749577	
7	Singapore Dollar	1.366872	0.731597	
8	Swiss Franc	0.995662	1.004357	
9	Malaysian Ringgit	4.158503	0.240471	
10	Japanese Yen	108.522481	0.009215	
11	Chinese Yuan Renminbi	6.917778	0.144555	
12				

Now let's walk through each of the icons in the integration flow and explain their purpose:

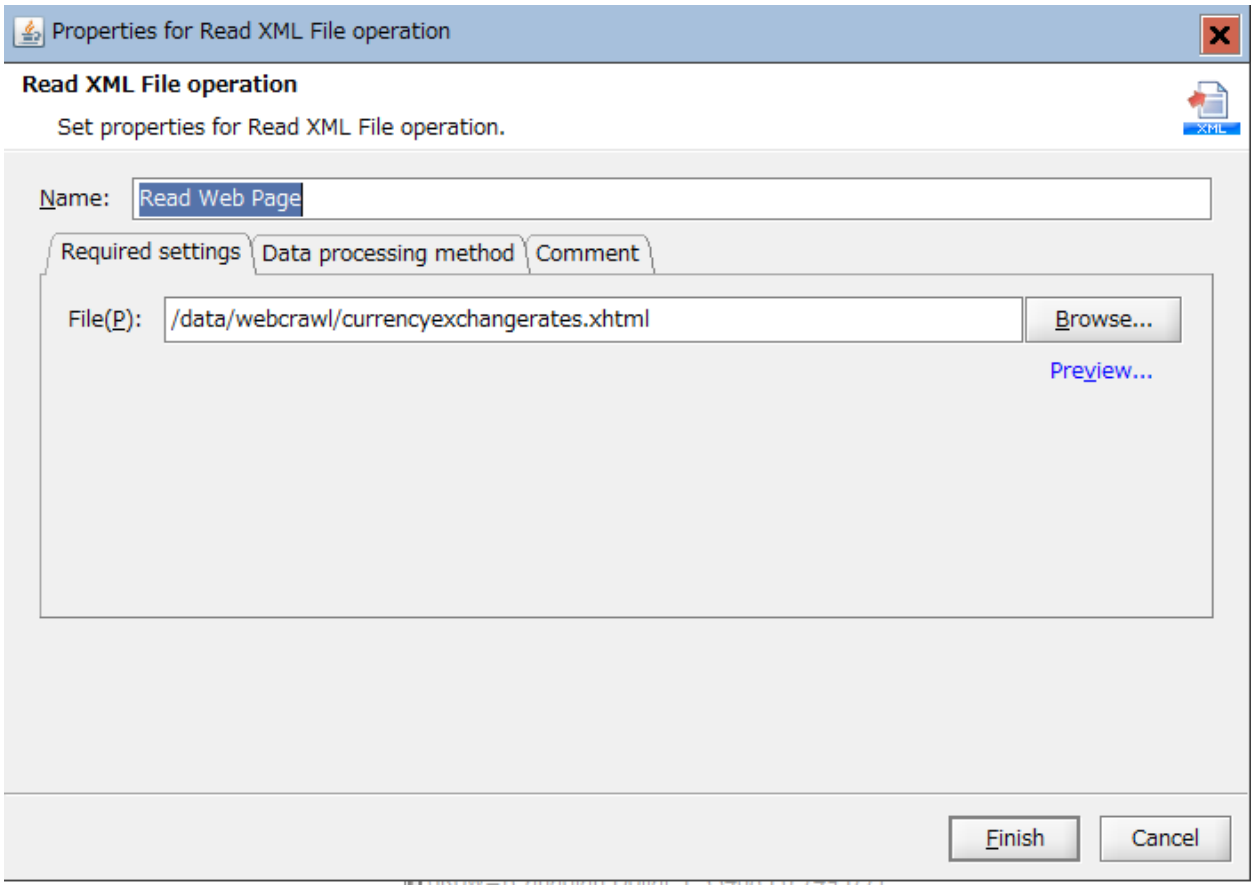
1. Store Web Page (Crawl adapter)

This adapter connects to the specific web URL to retrieve the web page and store it into a local file



2. Read Web Page (XML File Read adapter)

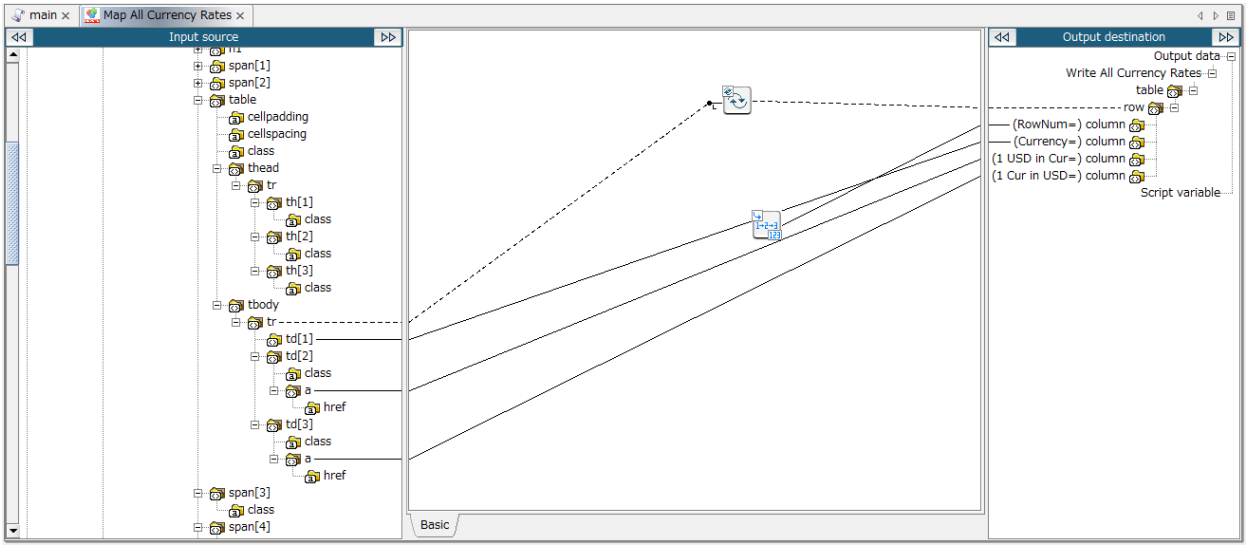
This adapter allows us to read the file created by the previous step



3. Map All Currency Rates

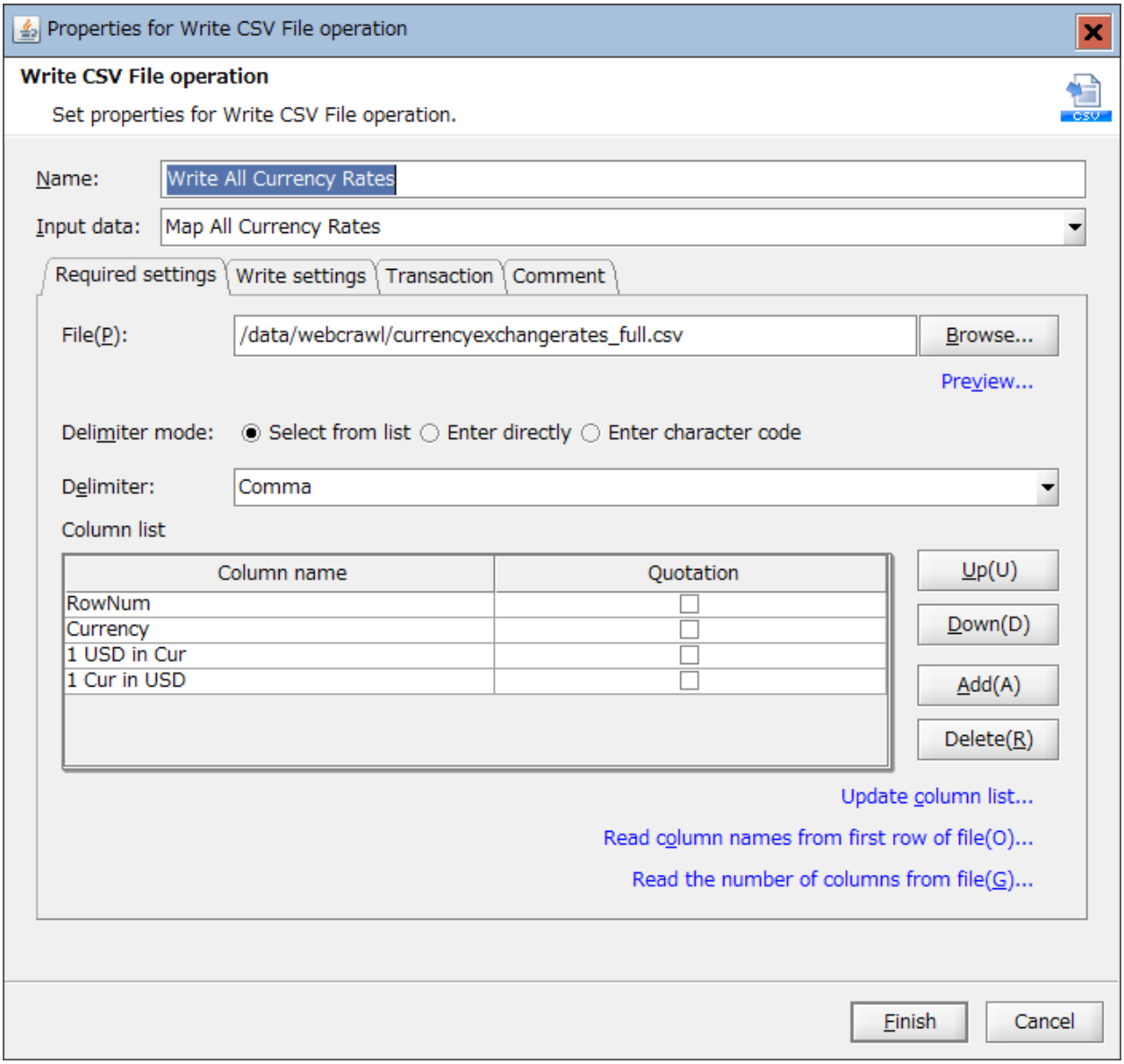
This component allows us to extract the precise fields, and nothing more, from the web page XML file and map to the structure of a CSV output file. This requires a hierarchical structure that's mapped to a relational structure, and shows the power of the mapping component. The schema for the hierarchical structure is automatically created by using the Load Schema feature of the mapper to read the web page in XML format and generate a corresponding schema. Another very powerful feature of HULFT Integrate's Mapper feature.

Due to the nature of the web page design, there is one table representing the top 10 exchange rates plus all the other exchange rates. At this point we capture all the rates and tag each output row with an incrementing row number. Later in the process we reduce the final output to just the top 10 rates (the first 10 rows).



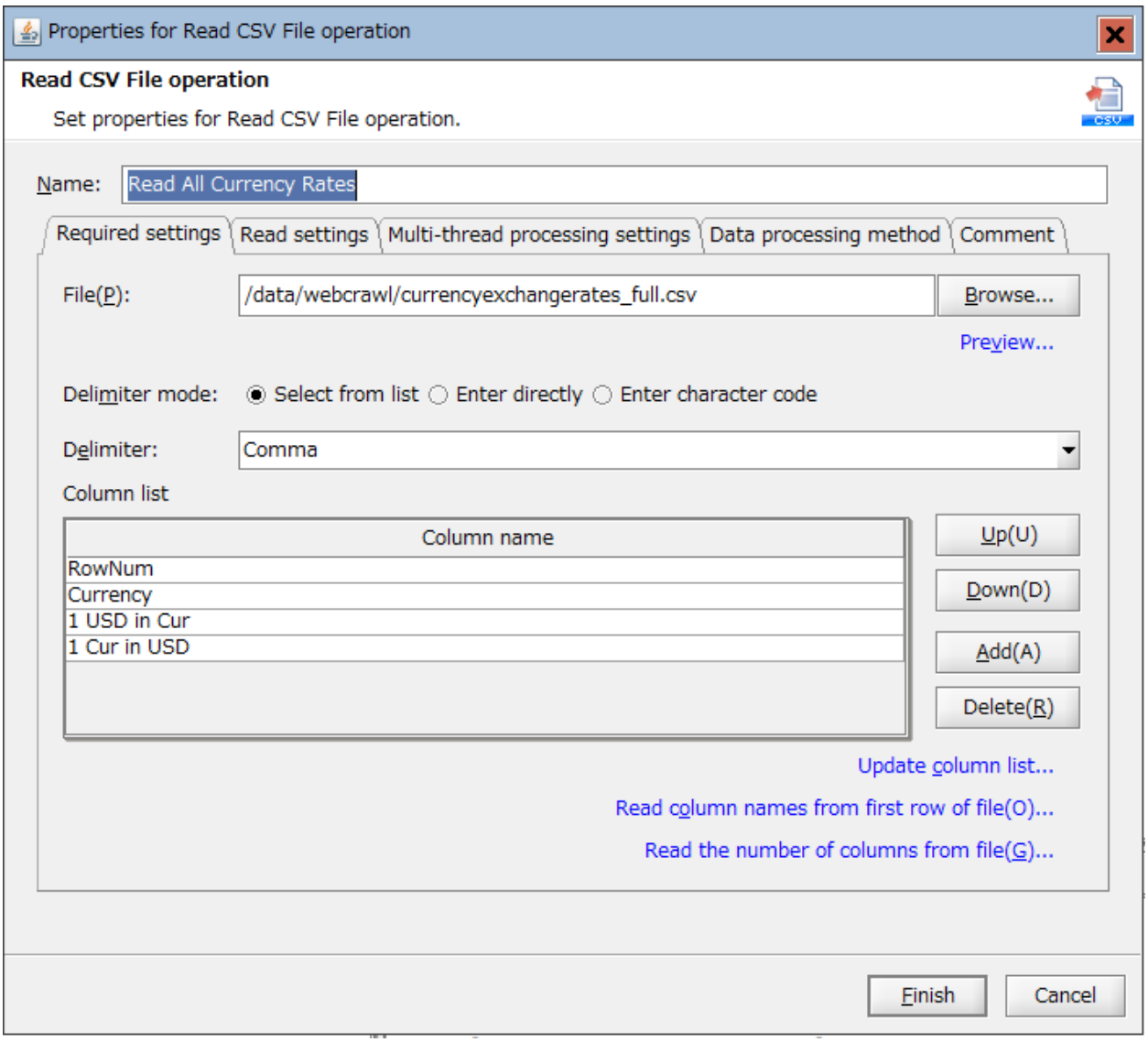
4. Write All Currency Rates (CSV File Write adapter)

Here we take the output from the previous Mapper component and write to a local CSV file.



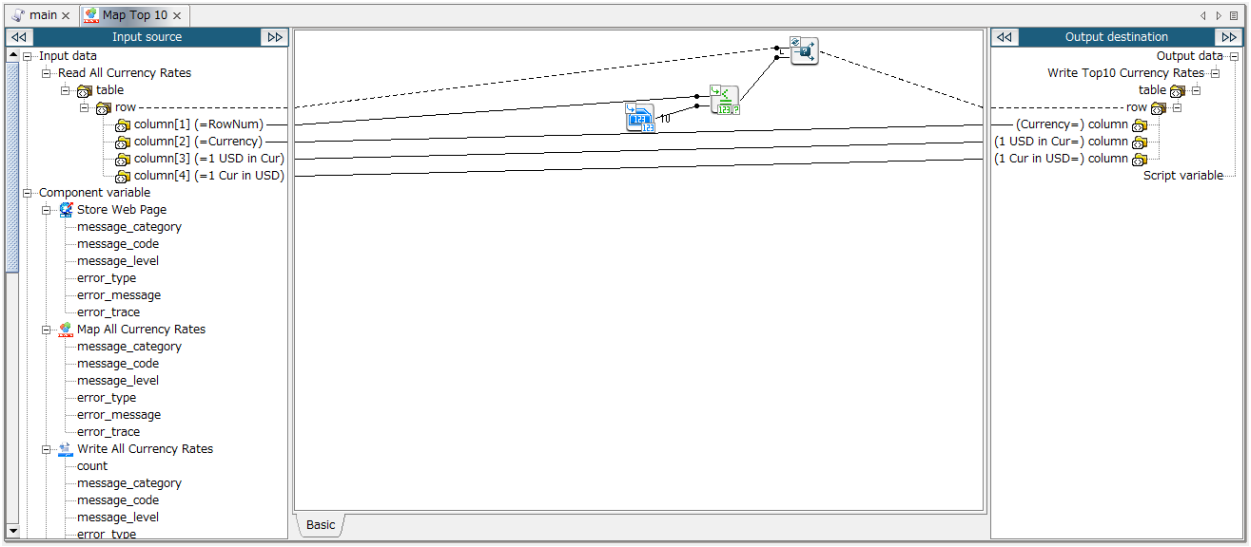
5. Read All Currency Rates (CSV File Read adapter)

This adapter reads all rows from the previously created CSV file containing all exchange rates



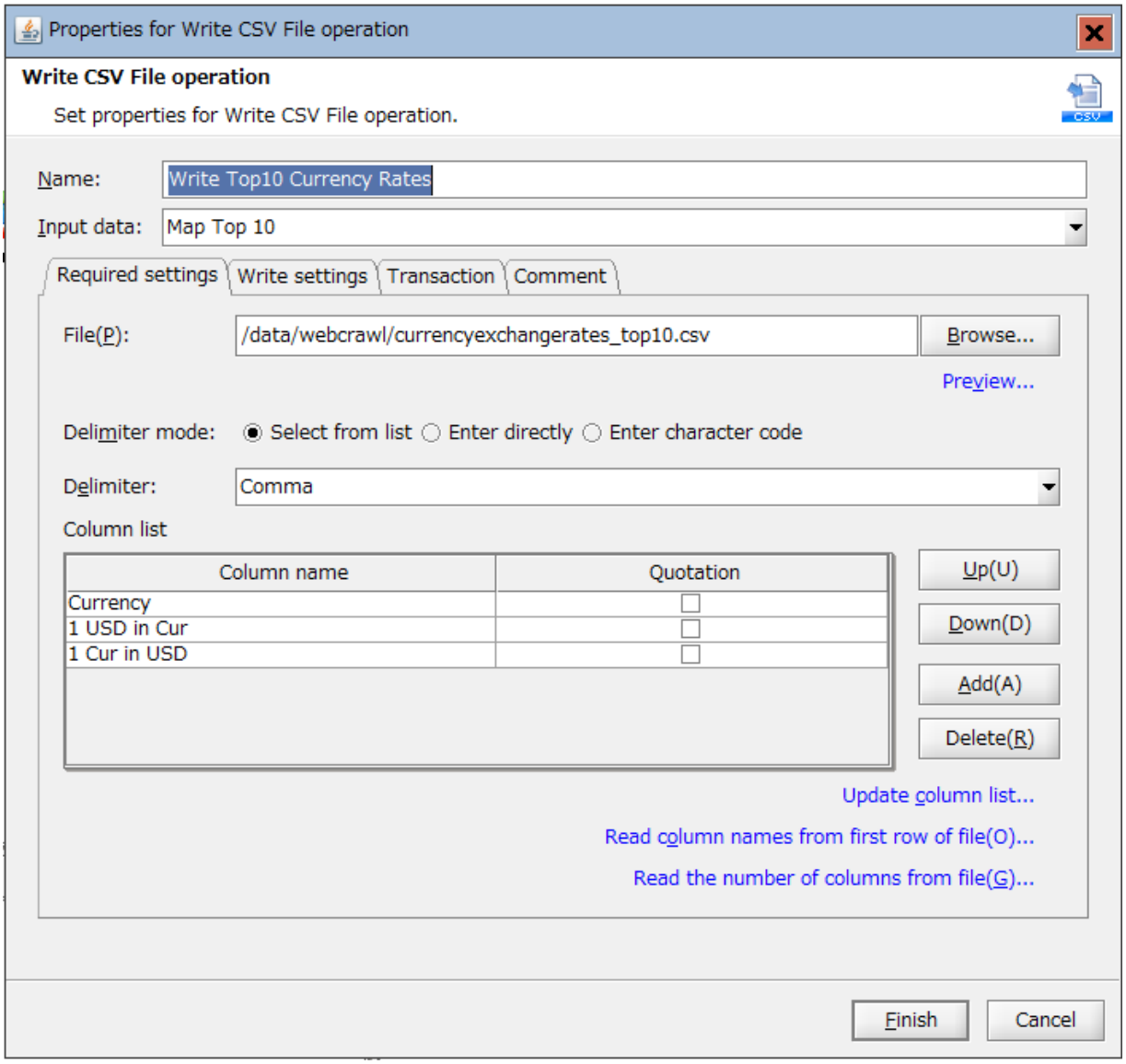
6. Map Top 10 (Mapper)

This component filters down the number of rows by only processing the first 10 rows and preparing that data for storage



7. Write Top 10 Currency Rates (CSV File Write adapter)

Takes the output from the previous Mapper and stores the exchange rate data into the final output file.



Summary

HULFT Integrate provides flexibility for your data integration problems. As we have demonstrated in this tutorial, some use cases may not seem to be an immediately obvious choice, but HULFT integrate is adaptable to a variety of uses and data sources.

Here we showed how HULFT Integrate can be used to accomplish screen scraping, for a case where there were no other legitimate means of extracting data in the short term. As I mentioned, screen scraping is an inadequate choice for long term integration needs due to its inherent fragility. There are always exceptions of course. For example, when screen scraping is a short term solution until a more viable long term option is available. In that case HULFT Integrate can be used to accomplish both the immediate and long term needs for integration.

Other Resources:

HULFT Integrate product sheet.

HULFT Integrate connects to all your diverse data sources, no matter where they reside. Here's a list of our adapters.

More than 10,000 customers across 43 countries trust HULFT. Learn more about our customers here.

Want to learn more?


Visit us at
https://hulftinc.com


Call us at
855-815-1518


Email us at
salesop@hulftinc.com