Introduction:

My motivation in writing this document was to provide beginner to intermediate level tips on managing those manual data conversion / transformation tasks that are usually managed using Excel. It is not unusual for these tasks to expand and take over more time than was budgeted at the time the task was identified.

The problem:

The story starts with a ERP implementation project and invariably includes that other system that was not scoped out because the e-commerce processes were deemed complex to build in the first roll out of the global solution.

There is an example of a file which provides an e-commerce system that needs the mapping of the Legacy customer numbers to the ERP customer numbers in the future state. This information will be stored in a table on the e-commerce server and is needed because e-commerce processes were not implemented on the ERP system and it was not implemented on the ERP system because … you get the picture.

The folks on the Data team have managed to map Legacy customer data over to the ERPP system. The crew from the e-commerce side is interested in continuing the seamless experience to the customer and therefore need the map from the Legacy customer to the SAP customer. This information is then assigned to the web user so that any given user can enter orders for their organization.

Table 1 File header for e-commerce system

UserName	CurrBillTo	CurrShipTo	SoldTo	ShipTo	BillTo	Payer
Jim_Bob_ Davis	894567	0	Blank	Blank	Blank	blank

Table 2 File header for customer data from ERP system

CNUM	NAME	AGRP	SGRP	PDIV	CHL	LNAME	LNUM	LDIV	YY_CCDB_NUM	CRDAT
205001	HALLS STORE LLC	ZPST	2223	10	10	HALLS STORE LLC	1013510	94	101351	

So far it looks simple enough for a consultant to pull two Excel files at the least and start putting together pivot tables and other such to compile an e-commerce map that has 132,000 entries on it. Alas, if it were that simple it would be the last thing the consultant would do before closing shop on Friday. So what makes this problem interesting and questionable as a manual task? I will get to that in short order.

To begin with the customer number on the e-commerce side is one or more of the 'CurrBillTo' or 'CurrShipTo' columns. For example, a customer that embodies a buying role, a ship to location and a payer role would be based off the 'CurrBillTo' customer. However, many of the retail locations are managed as ship to customer locations and these are considered a combination of a 'Bill-To' number and an increment in the 'Ship-To' number. So the Legacy customer number is really '8945670' if it is a 'complete' customer id or it is for example '894567100' if it is the 100th retail location for this customer. Further a 'complete' customer that embodies the sold-to, ship-to, bill-to and payer roles will have the

same ERP customer number for all of the roles. On the other hand, ship-to only locations will be maintained as 'ship-to' customers in the ERP system. You are probably starting to think of joins already. Consider that we have two Excel csv files that we start with.

Is it likely that when all is said and done more rules on matching and joining the data sets would have been deliberated, processed and implemented. So what is a good tool? There are many open source tools and licensed ones in the market. I will be explaining the approach using an open source tool, it was open source at the time I prepared the document.

The tool I was chosen for this exercise is 'KNIME'. KNIME stands for **Kon**stanz **I**nformation **M**in**E**r and is pronounced: [naim] (that is, with a silent "k", just as in "knife"). It is developed by KNIME.com AG located in Zurich and the group of Michael Berthold at the University of Konstanz, Chair for Bioinformatics and Information Mining. According to Wikipedia from the web page 'https://en.wikipedia.org/wiki/KNIME', it is an open source data analytics, reporting and integration platform. A graphical user interface allows assembly of nodes for data preprocessing (ETL: Extraction, Transformation, Loading), for modeling and data analysis and visualization.

KNIME allows users to visually create data flows (or pipelines), selectively execute some or all analysis steps, and later inspect the results, models, and interactive views. These are also called workflows.

This document is not an exposition of the KNIME tool, instead I will focus on how a workflow can be setup for a simple to medium complex use case.

Workflow / Data Pipeline:

The following flow diagram is the pipeline that I have created to process the use case.

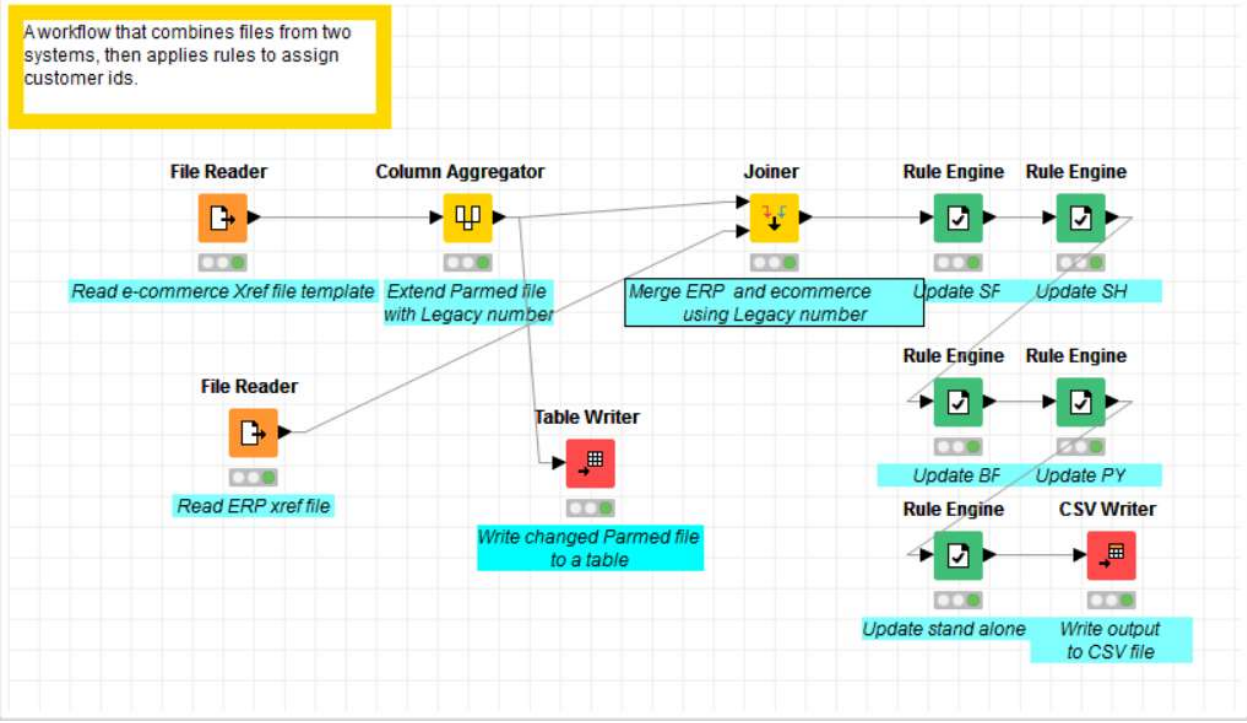


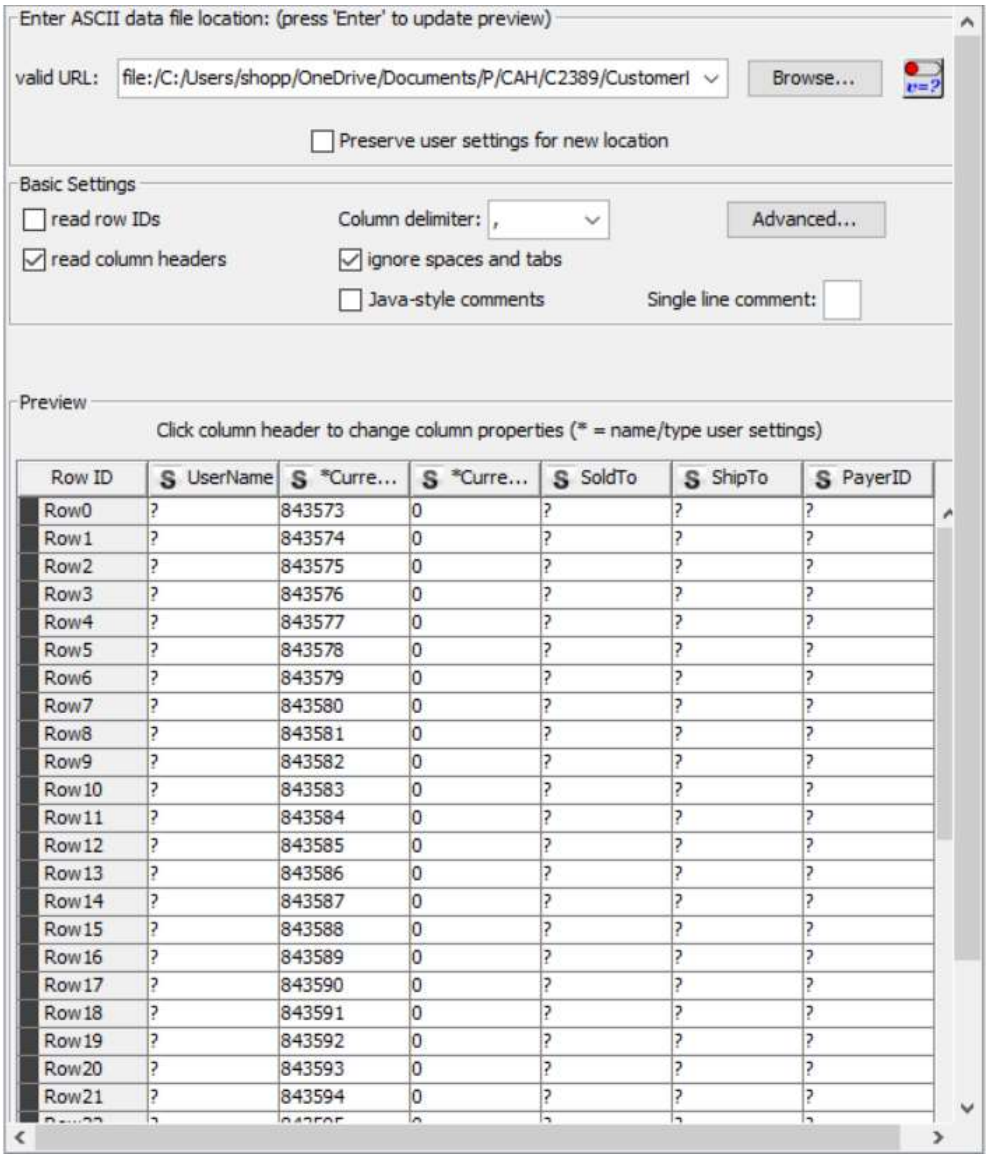
Figure 1 Data pipeline to generate cross reference file automatically

I will go through the various elements of this workflow next.

FileReader

FileReader will read in data from an ASCII file or a URL. Various typical options as well as more advanced ones are available to drive the storage of data.

Here I read in the comma delimited data with a header.



Row ID	UserName	*Curre...	*Curre...	SoldTo	ShipTo	PayerID
Row0	?	843573	0	?	?	?
Row1	?	843574	0	?	?	?
Row2	?	843575	0	?	?	?
Row3	?	843576	0	?	?	?
Row4	?	843577	0	?	?	?
Row5	?	843578	0	?	?	?
Row6	?	843579	0	?	?	?
Row7	?	843580	0	?	?	?
Row8	?	843581	0	?	?	?
Row9	?	843582	0	?	?	?
Row10	?	843583	0	?	?	?
Row11	?	843584	0	?	?	?
Row12	?	843585	0	?	?	?
Row13	?	843586	0	?	?	?
Row14	?	843587	0	?	?	?
Row15	?	843588	0	?	?	?
Row16	?	843589	0	?	?	?
Row17	?	843590	0	?	?	?
Row18	?	843591	0	?	?	?
Row19	?	843592	0	?	?	?
Row20	?	843593	0	?	?	?
Row21	?	843594	0	?	?	?

Figure 2 Configuring the file reader

Various options are available such as quote control, missing value mapping, decimal value separators and so on. The result of this process is that data is read from the file into memory. Another file reader is used to store the ERP customer data into memory.

Column Aggregator

One of the transformations that we have to do with the e-commerce data is the generation of the Legacy customer number. For this we use the *Column Aggregator.* The aggregator combines the columns holding the Bill to and Ship to numbers to generate the Legacy customer number column.

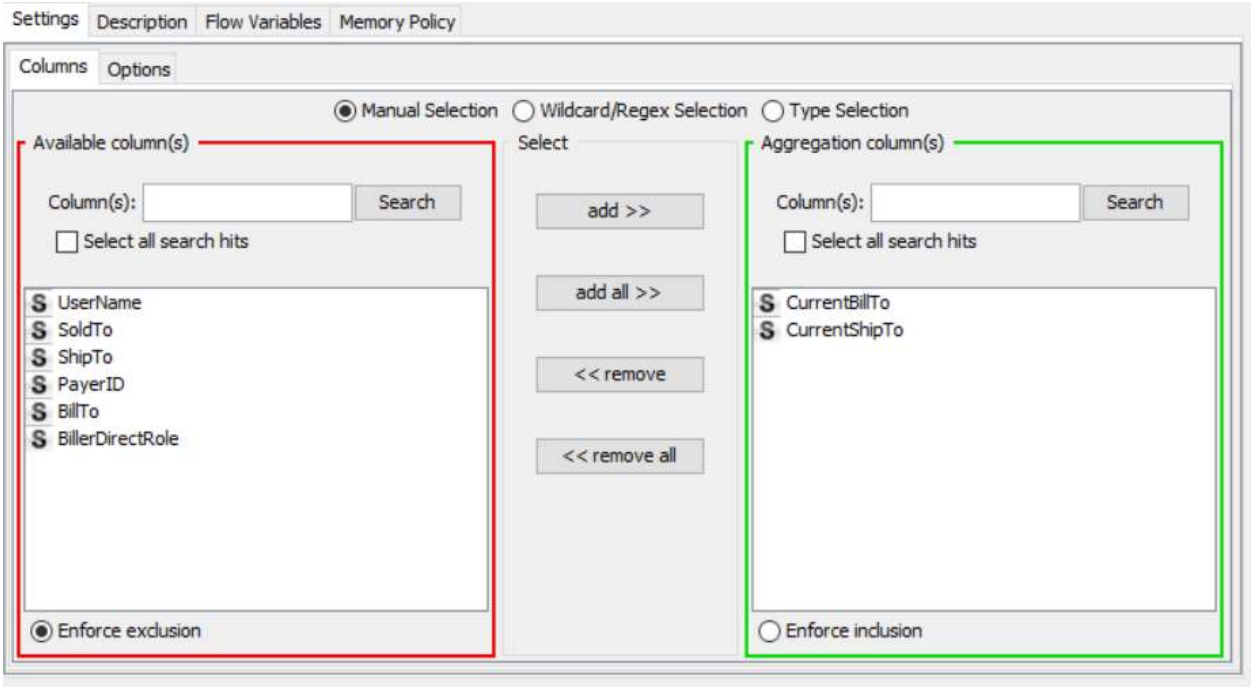


Figure 3 Column Aggregation to generate Legacy customer numbers

Here is an excerpt of the output.

Row ID	UserName	Current...	Current...	SoldTo	ShipTo	PayerID	BillTo	BillerDir...	YY_LEG...
Row0	?	843573	0	?	?	?	?	?	8435730
Row1	?	843574	0	?	?	?	?	?	8435740
Row2	?	843575	0	?	?	?	?	?	8435750
Row3	?	843576	0	?	?	?	?	?	8435760
Row4	?	843577	0	?	?	?	?	?	8435770
Row5	?	843578	0	?	?	?	?	?	8435780
Row6	?	843579	0	?	?	?	?	?	8435790

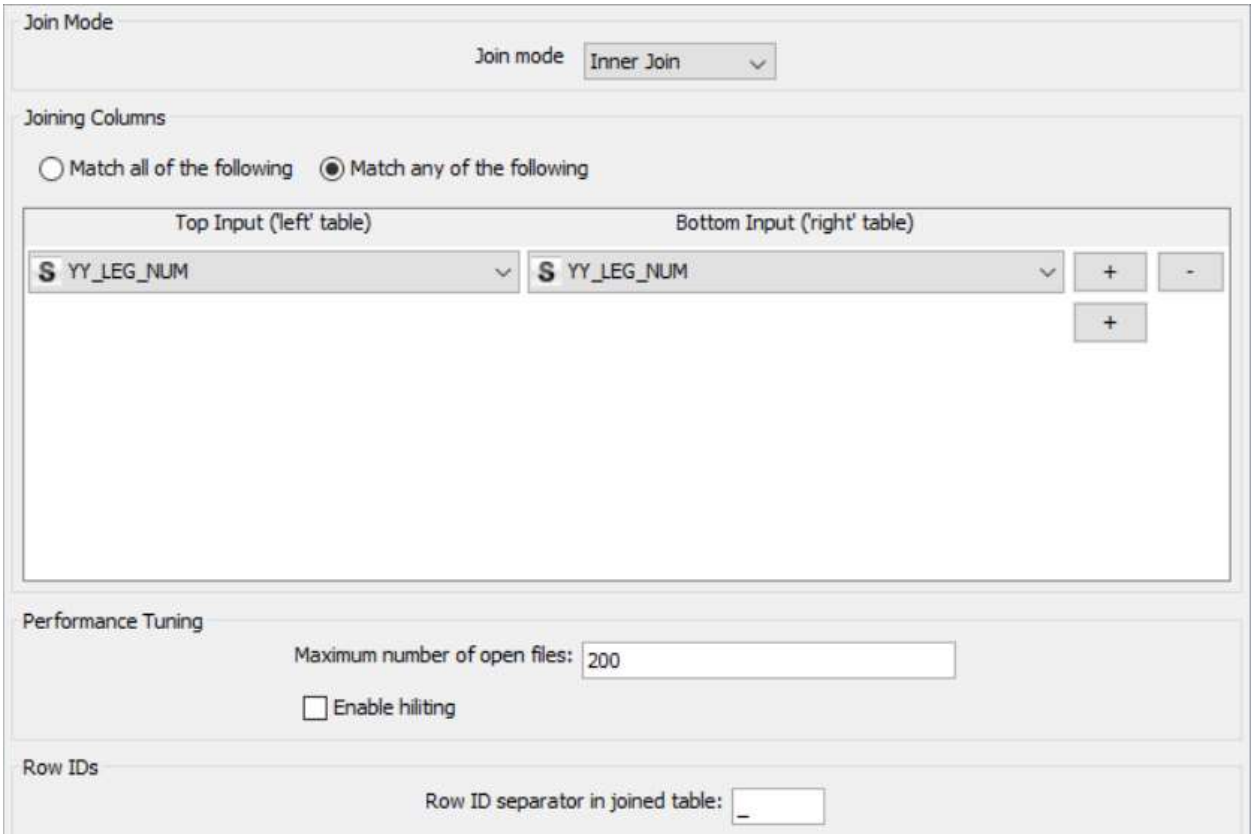
Figure 4 Excerpt of the Column Aggregator process

The utility of these pipeline tasks lies in the fact that each process holds its own copy of the operation that it was tasked with. So, if you need to change a subsequent process you do not typically need to change upstream steps unless the data model or rules have to be changed.

Joiner

The next step in the process is to consolidate the data or merge the data from the e-commerce and erp data sets. We use the *Joiner* step for this purpose. The join or merge process is based on columns in both data sets.

Various join options are available; I will be using the 'inner join'. Admittedly this is not the best choice, but keep in mind that I have created an example. In reality one would need to possibly consider a 'left-outer' join. I leave it to the readers to evaluate the options.



The result of this operation generates a data set like the one below.

Row ID	UserName	Current...	Current...	SoldTo	ShipTo	PayerID	BillTo	BillerDir...	YY_LEG...	YY_KU...	YY_NA...
Row3_Row3382	?	843576	0	?	?	?	?	?	8435760	2050003367	JEMAS LLC
Row6_Row6930	?	843579	0	?	?	?	?	?	8435790	2050006923	THE MEDICI...
Row36_Row6...	?	843609	0	?	?	?	?	?	8436090	2050006969	THE PEOPLE...
Row47_Row7...	?	843620	0	?	?	?	?	?	8436200	2050007034	THRIFTEE S...
Row57_Row7...	?	843630	0	?	?	?	?	?	8436300	2050007082	TOP CARE P...
Row58_Row7...	?	843631	0	?	?	?	?	?	8436310	2050007089	TOTAL CAR...
Row62_Row7...	?	843635	0	?	?	?	?	?	8436350	2050007098	TOWER PHA...
Row72_Row7...	?	843645	0	?	?	?	?	?	8436450	2050007135	TRI-C MEDI...
Row74_Row7...	?	843647	0	?	?	?	?	?	8436470	2050007147	TRINITY PH...
Row77_Row7...	?	843650	0	?	?	?	?	?	8436500	2050007152	TRINITY RX

Figure 5 Joined data set based on Legacy customer number

There is a neat little trick that KNIME uses to identify the rows from the two data sets that were merged. See if you can locate them.

Let us take quick stock of where we are. We have managed to merge the two data sets. Good work. What remains to be done in our example is to transfer the ERP customer number in the column 'YY_KUNNR' to the Sold-to, Ship-to, Bill-to and Payer columns based on the customer role. The customer role is defined by values in the column 'YY_KTOKD'.

Rules Engine

In this *Rules Engine* step, we create rules that are used by the tool to evaluate every row from the joined dataset and make changes to the corresponding column values.

Here is an example of a rule that updates the 'Sold to' column. The rule bases its update on the value of the customer role grouping from the ERP system. If there is a match the value of the ERP customer is transferred to the 'Sold-to' column of that row.

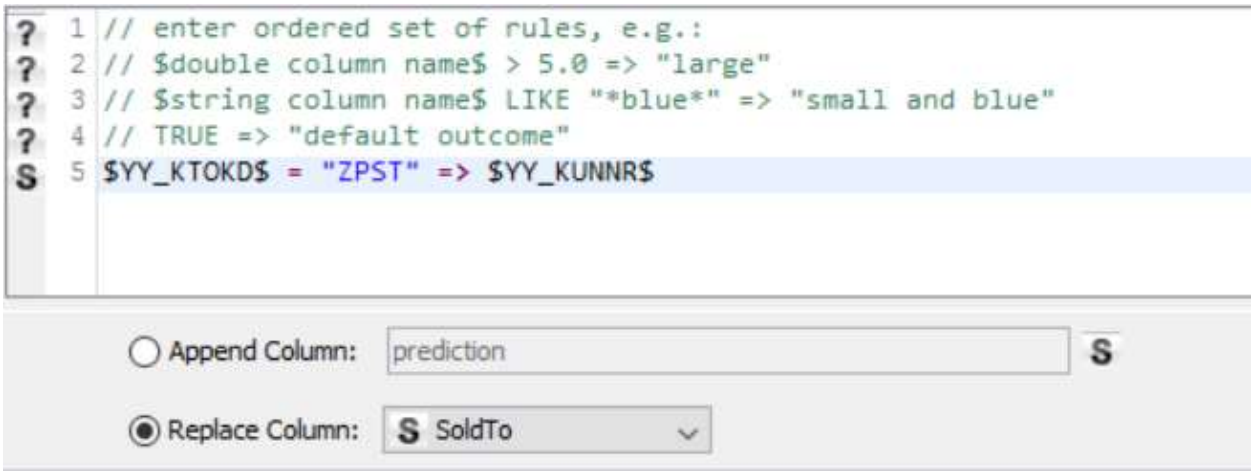


Figure 6 Rule to update Sold-to column

Here is an example of the results.

Row ID	UserName	Current...	Current...	SoldTo	ShipTo	PayerID	BillTo	BillerDir...	YY_LEG...	YY_KU...
Row3_Row3382	?	843576	0	2050003367	?	?	?	?	8435760	2050003367
Row6_Row6930	?	843579	0	2050006923	?	?	?	?	8435790	2050006923
Row36_Row6...	?	843609	0	2050006969	?	?	?	?	8436090	2050006969

Figure 7 Example of updated Sold-to column

Similar rules are defined to update the Ship-to, Bill-to and Payer columns for the 'Sold-to' customer. A separate rule is defined to update the 'Ship-to' column only for 'Ship-to' customers.

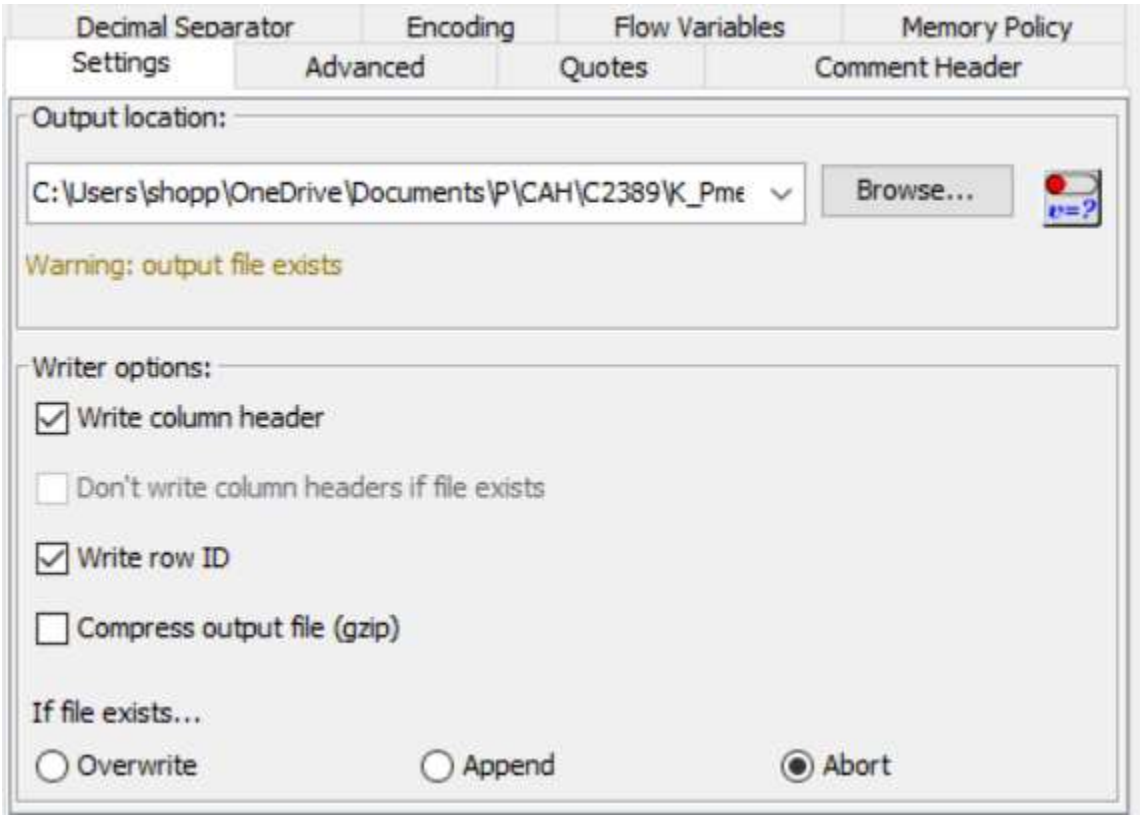
The end result looks like the following list.

Row ID	UserName	Current...	Current...	SoldTo	ShipTo	PayerID	BillTo
Row3_Row3382	?	843576	0	2050003367	2050003367	2050003367	2050003367
Row6_Row6930	?	843579	0	2050006923	2050006923	2050006923	2050006923
Row36_Row6...	?	843609	0	2050006969	2050006969	2050006969	2050006969
Row47_Row7...	?	843620	0	2050007034	2050007034	2050007034	2050007034
Row57_Row7...	?	843630	0	2050007082	2050007082	2050007082	2050007082
Row58_Row7...	?	843631	0	2050007089	2050007089	2050007089	2050007089
Row62_Row7...	?	843635	0	2050007098	2050007098	2050007098	2050007098

Figure 8 Updated dataset

CSV Writer

Finally, we would like to create another csv file to pull the merged updated dataset and provide it to the e-commerce team for loading into the e-commerce server. The *CSV Writer* is used to complete this task.



Various options are available to customize the output file.

Concluding remarks

I have used production scale data sets in this example, admittedly your production scale need not be the same as mine. Let me provide some sense of the scale here. The e-commerce file had 130,000 rows. The ERP file only had 9000 records. The workflow performed without hiccups. I would recommend some solid validation to confirm robustness of the transformation. What I had set out to do was to show an easy way to handle tasks that could easily end up consuming ten times more time than originally budgeted. This workflow is scalable and can be replicated. My review of the task types available confirms the availability of many more kinds of tasks particularly ones of the input, output and transformation kind. I leave it to the user to explore more of the options.

References

@INPROCEEDINGS{BCDG+07,

author = {Michael R. Berthold and Nicolas Cebron and Fabian Dill and Thomas R. Gabriel and Tobias K\"{o}tter and Thorsten Meinl and Peter Ohl and Christoph Sieb and Kilian Thiel and Bernd Wiswedel},

title = {{KNIME}: The {K}onstanz {I}nformation {M}iner},

booktitle = {Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)},

publisher = {Springer},

ISBN = {978-3-540-78239-1},

ISSN = {1431-8814},

year = {2007}

}

Footnotes¶

1) author = {Michael R. Berthold and Nicolas Cebron and Fabian Dill and Thomas R. Gabriel and Tobias K\"{o}tter and Thorsten Meinl and Peter Ohl and Christoph Sieb and Kilian Thiel and Bernd Wiswedel},

title = {{KNIME}: The {K}onstanz {I}nformation {M}iner},

booktitle = {Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)},

publisher = {Springer},

ISBN = {978-3-540-78239-1},

ISSN = {1431-8814},

year = {2007}

}