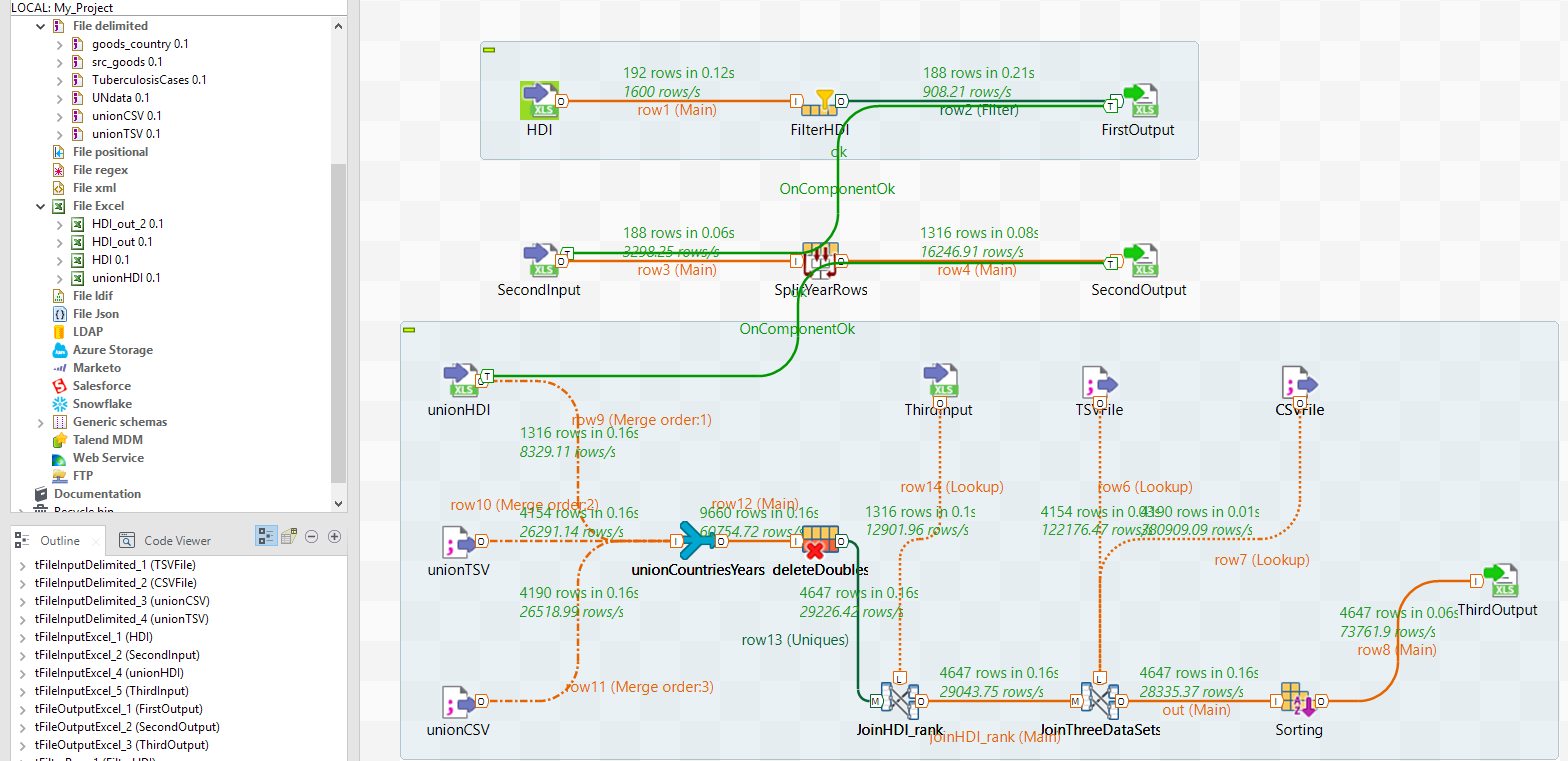
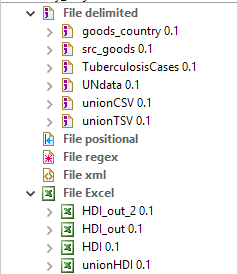
**Module1**

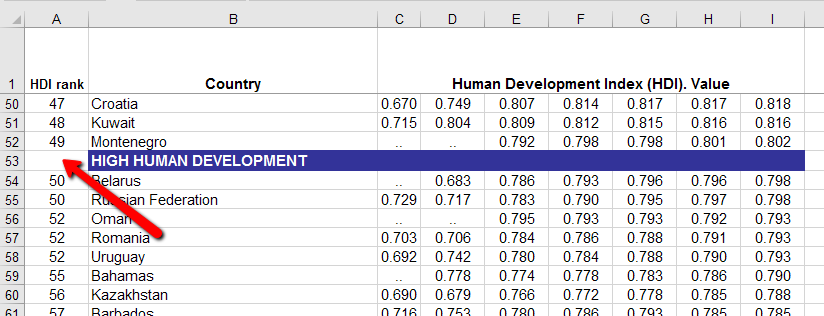


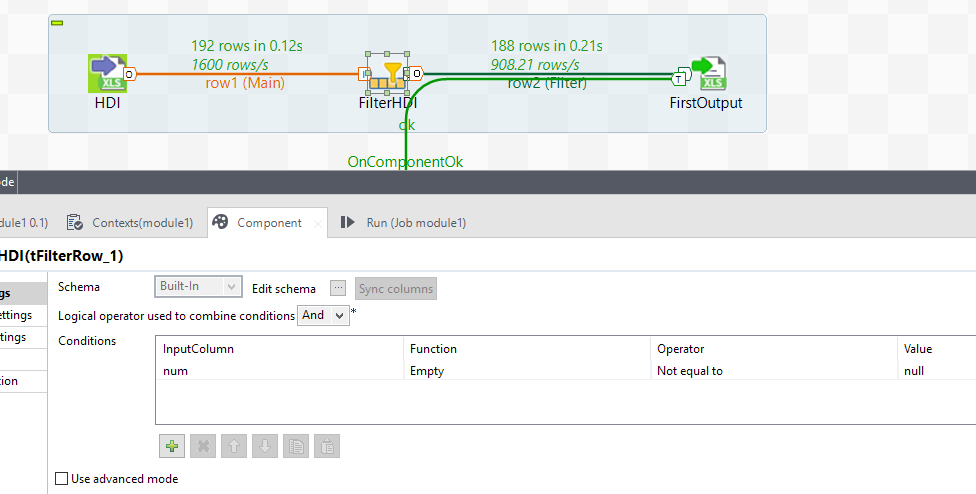
1. Create metadata files.



1. First subjob – Filter Human Development Index trends, 1990–2014.xls file.

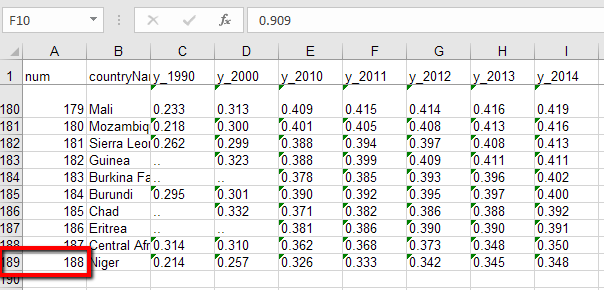
As you can see on the picture below the HDI Rank Column has gaps (null values) near by the Group delimiters. Use this information on the filter stage.





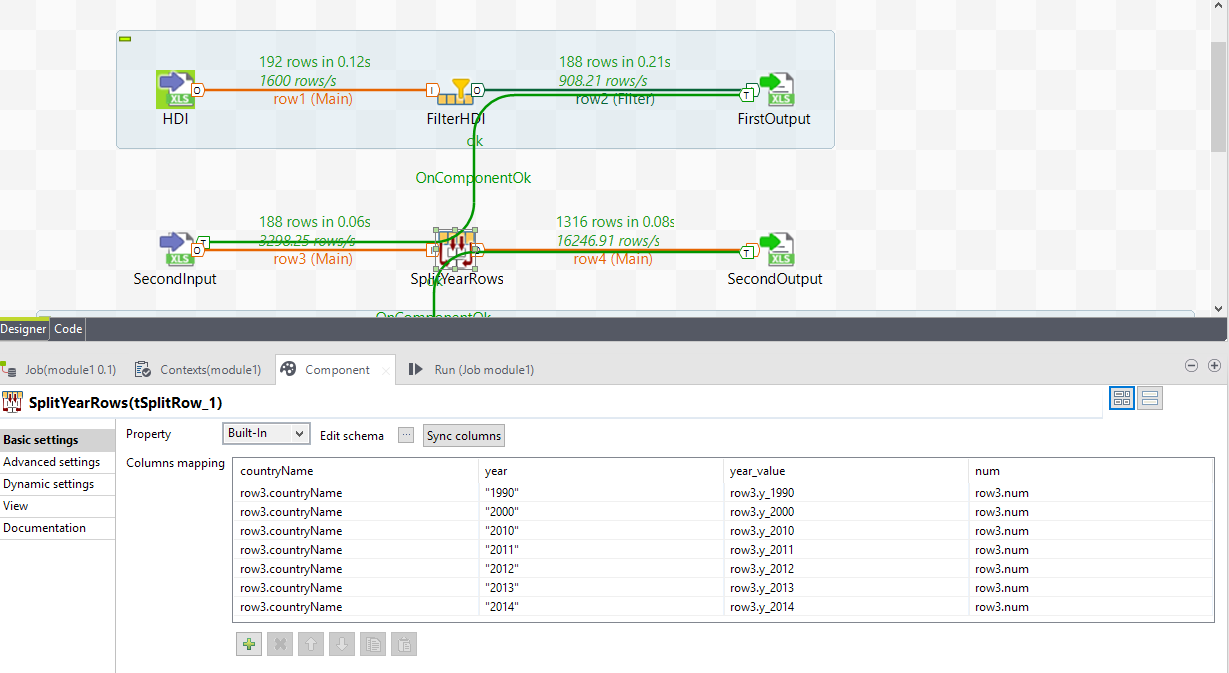
*To make results more obviously there have been created some OUT files after every step.*

First output file looks like on the picture below. Quantity of rows are the same (first line is Header).



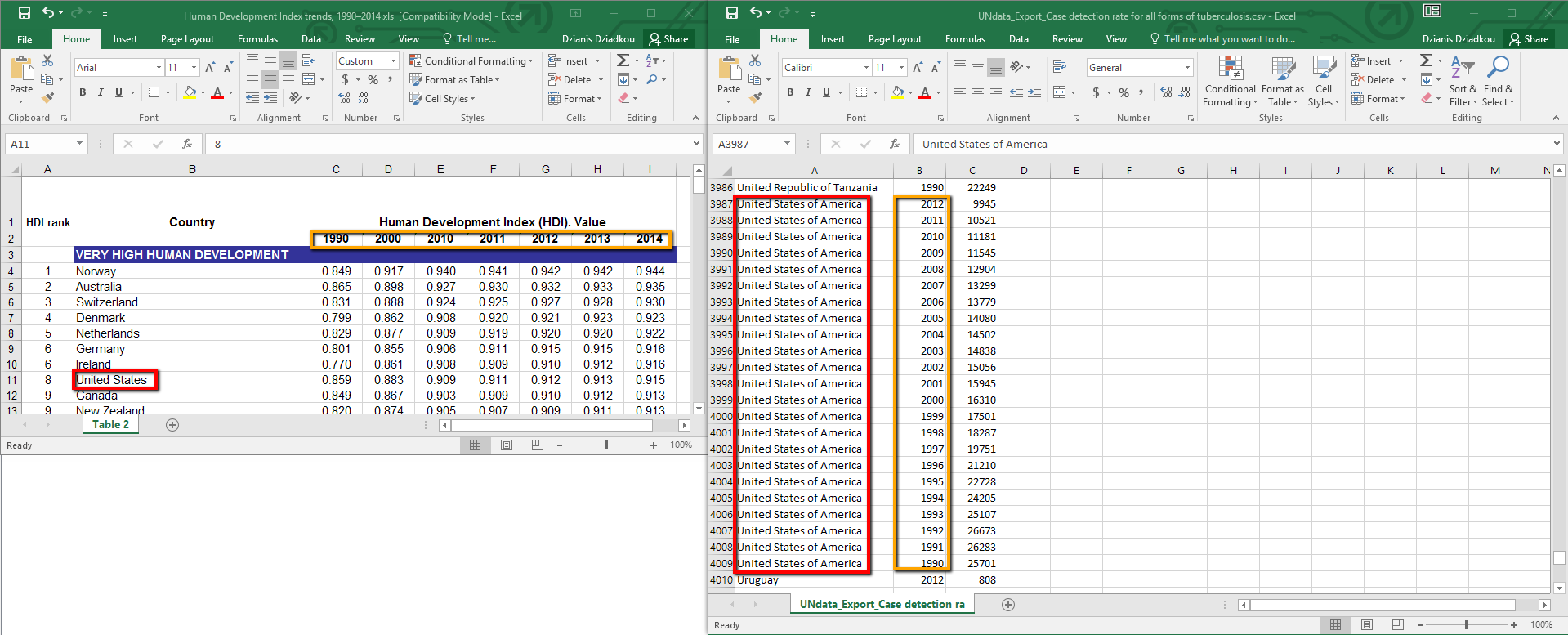
1. Unpivot data subjob.

Using split stage allowed unpivot the main table. ***YEAR*** values were added by hands.

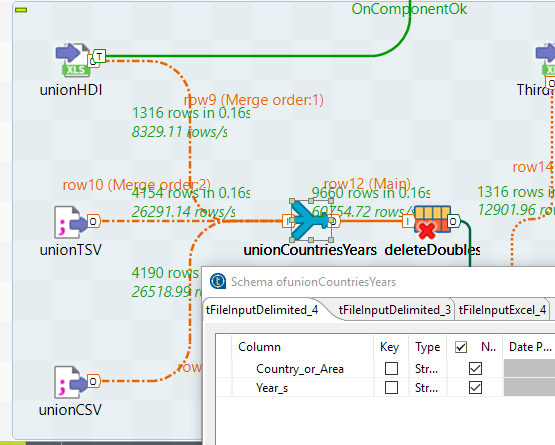


1. Join data sets subjob.

Some files contain such countries (red highlight) or years (orange highlight) which don’t exist in other files. Some of them even have different name.



To avoid creating gaps it is necessary to use UNION from three data sets to collect all possible years and countries.

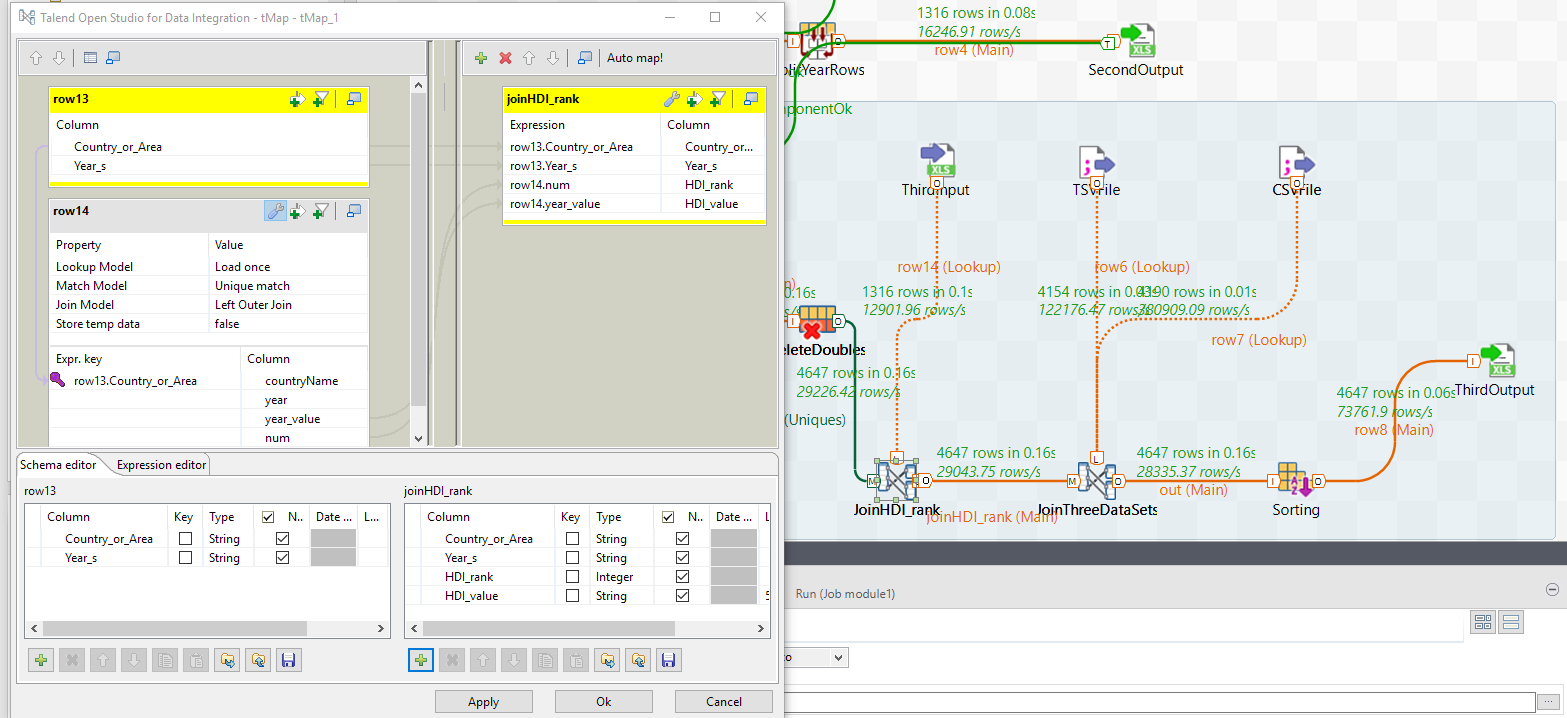


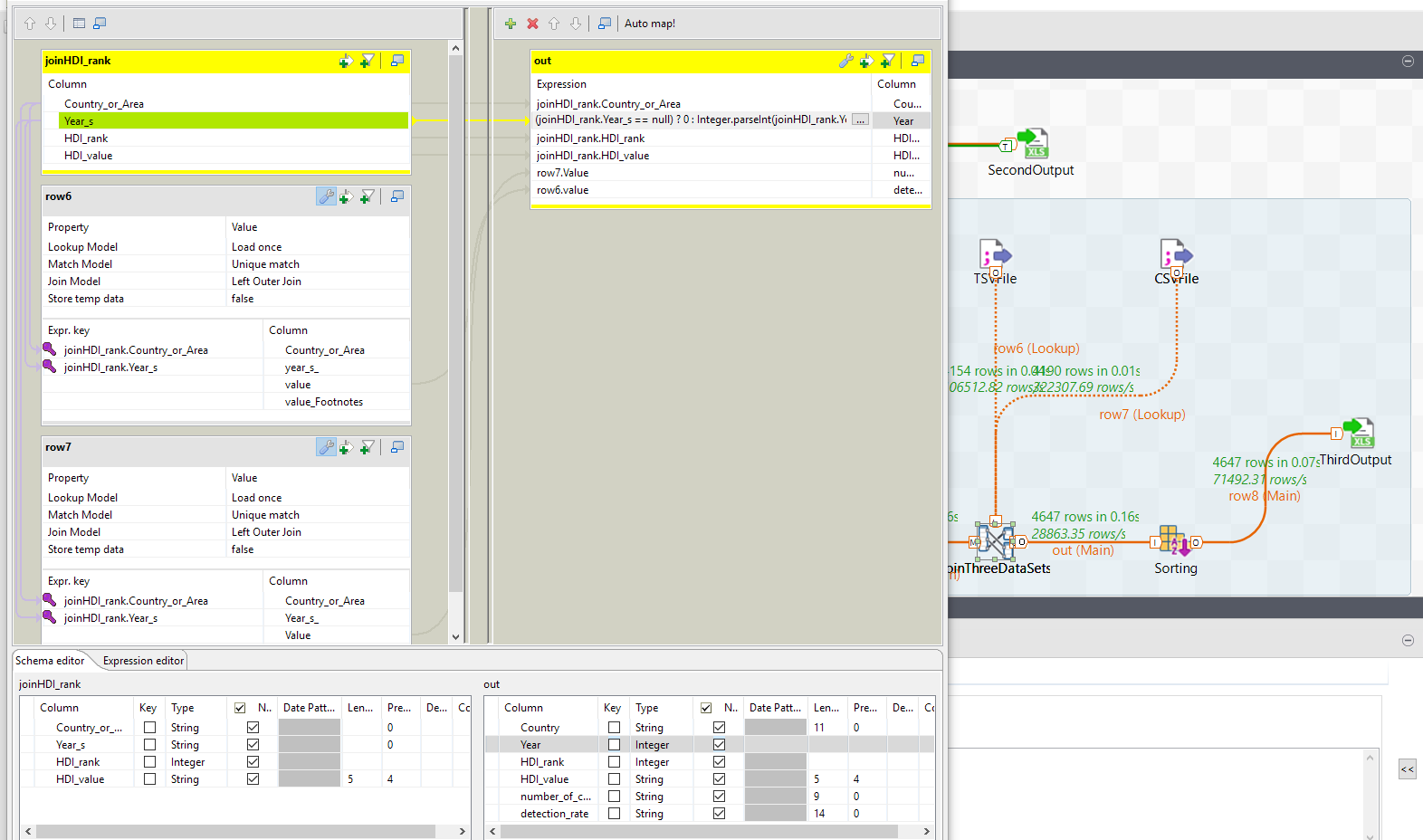
All files should have the same schemas and structures (column names and quantity of columns).

Then delete duplicate rows to make result set not redundant.

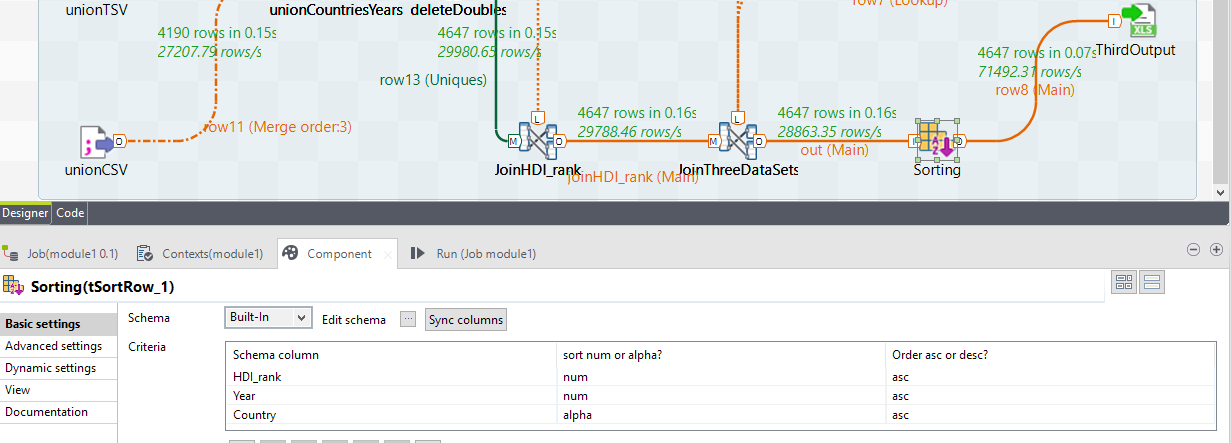
Regarding the task the ***HDI\_rank*** column should be presented in the result data set thus it’s necessary to use ***HDI\_rank*** (***num*** column in metadata) before the main join in this step or some ***HDI\_rank*** would be lost. For instance, for those countries which exist in one file and have different names in another. On the picture below, you can see that ***num*** column comes from another data set.

This way those countries which weren’t exist in Human Development Index trends, 1990–2014.xls file will not have ***HDI\_rank*** value.



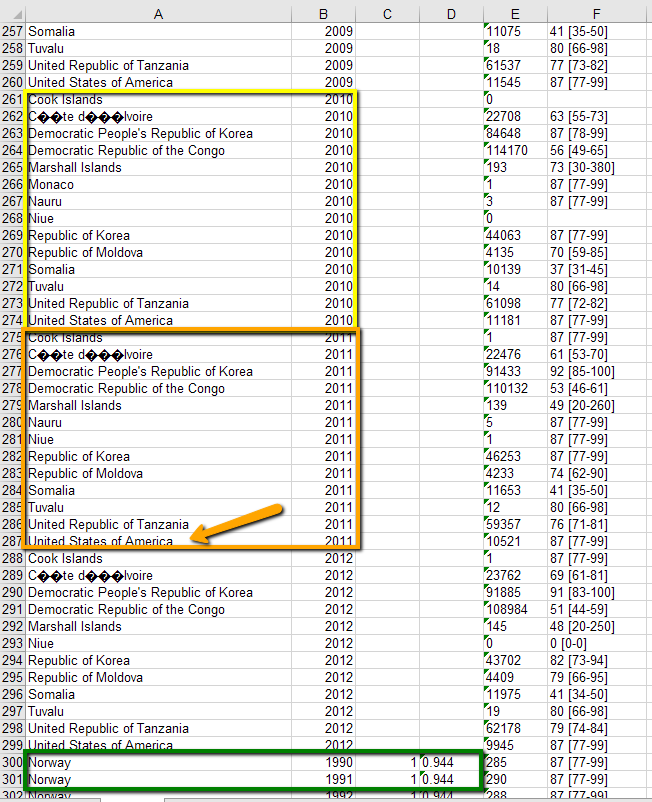


Next do the sorting stage. As there were no any requirements about types of orders (ASC/DESC) this stage could be realized as on the picture below.

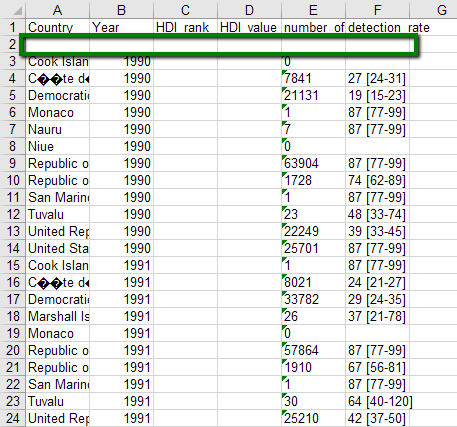


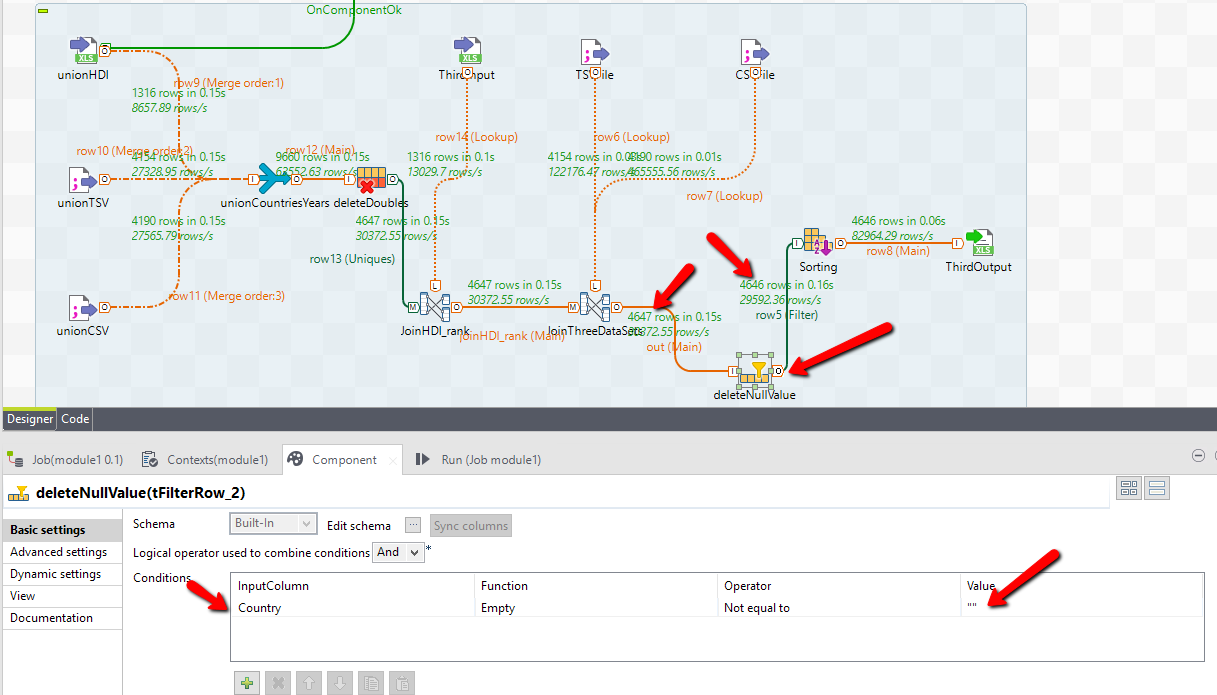
Final data set is on the picture below. Yellow and orange highlighted boxes have almost the same countries. Some of these countries didn’t repeat. As it has been mentioned before because countries have different names they weren’t joined -> don’t have ***HDI\_rank***.

Green box shows that rows ordered by ***HDI\_rank***, ***Year*** and then by ***Country*** columns.



To avoid empty strings in the output document coming from the join step (JoinThreeDataSets) we could add filter like on the picture below.





Because there was not a requirement about it I didn’t add this filter to the final project version.