**Practical No 2**

**DATA WAREHOUSING USING PENTAHO**

**Data Warehousing**

Data warehousing is the process of constructing and using a data warehouse. A data

warehouse is constructed by integrating data from multiple heterogeneous sources

that support analytical reporting, structured and/or ad hoc queries, and decision

making. Data warehousing involves data cleaning, data integration, and data

consolidations.

The following are the functions of data warehouse tools and utilities -

* **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
* **Data Cleaning** − Involves finding and correcting the errors in data.
* **Data Transformation** − Involves converting the data from legacy format to warehouse format.
* **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
* **Refreshing** − Involves updating from data sources to warehouse.

**Who needs Data warehouse?**

DWH (Data warehouse) is needed for all types of users like:

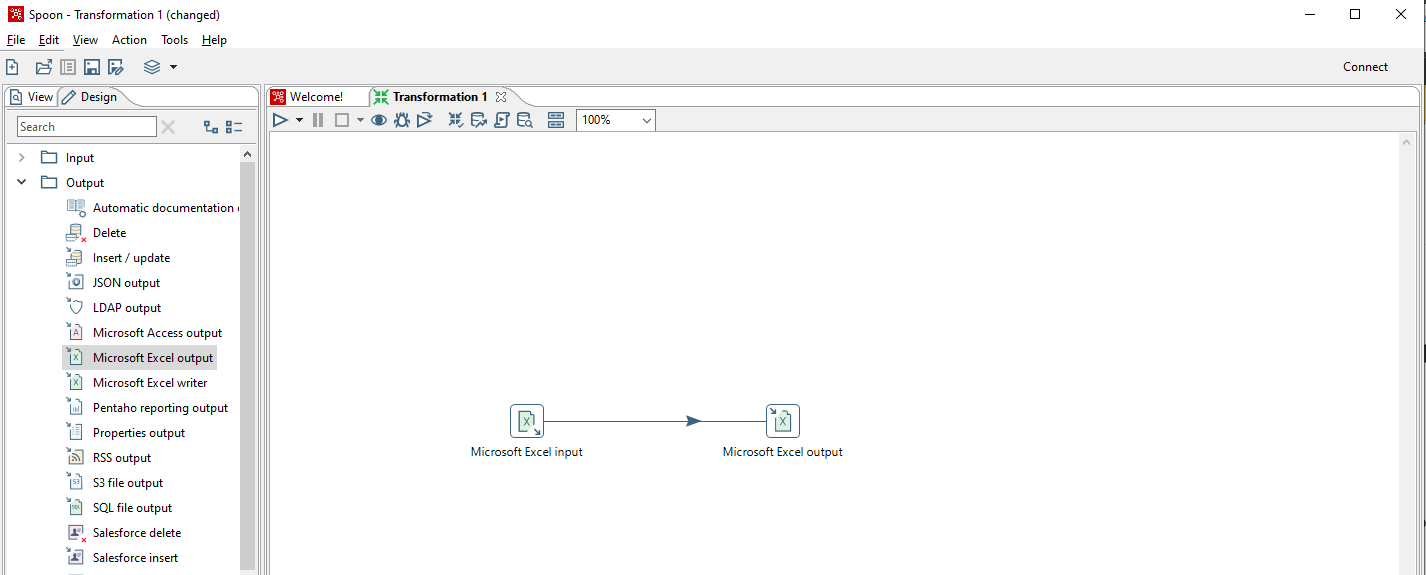
* Decision makers who rely on mass amount of data.
* Users who use customized, complex processes to obtain information from multiple data sources.
* It is also used by the people who want simple technology to access the data.
* It also essential for those people who want a systematic approach for making decisions.
* If the user wants fast performance on a huge amount of data which is a necessity for reports, grids or charts, then Data warehouse proves useful.
* Data warehouse is a first step If you want to discover 'hidden patterns' of data-flows and groupings.

**Advantages of Data Warehouse (DWH):**

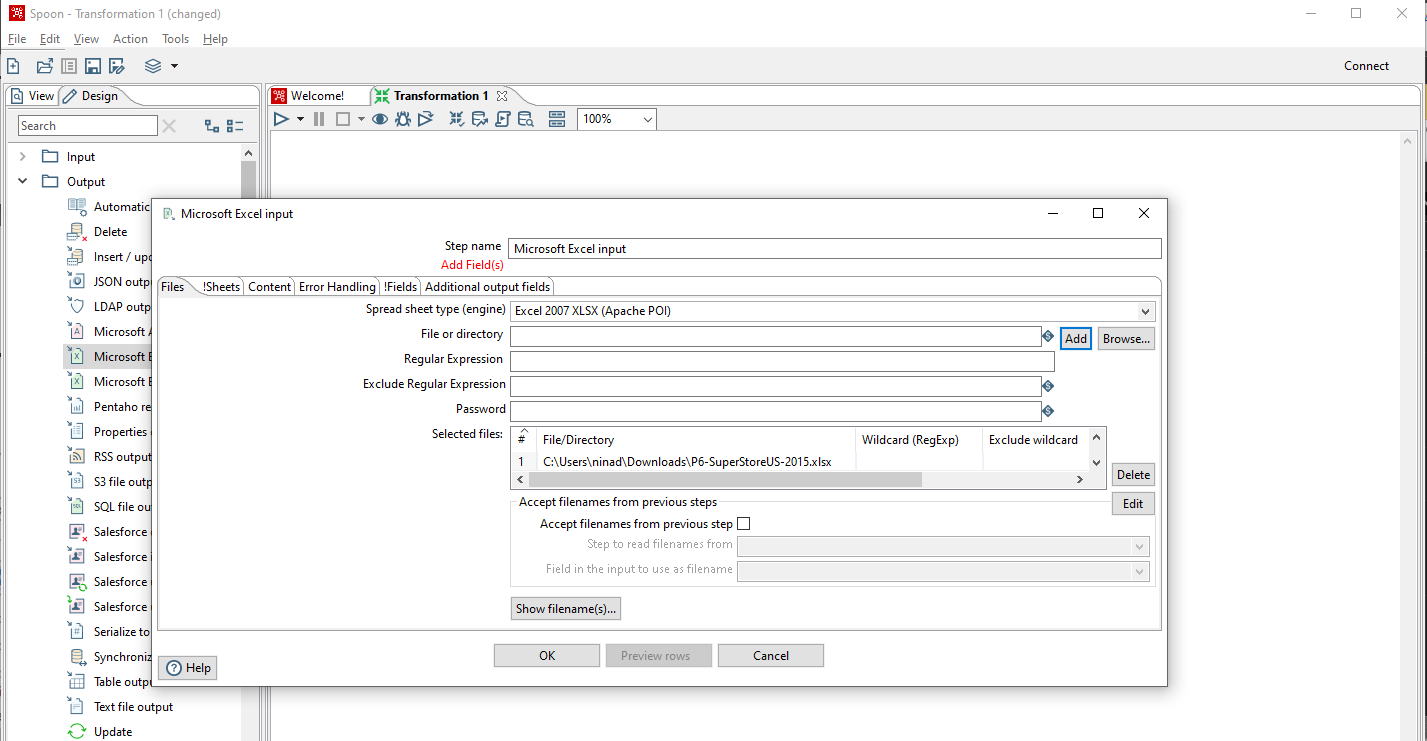
* Data warehouse allows business users to quickly access critical data from some sources all in one place.
* Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
* Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
* Data warehouse helps to reduce total turnaround time for analysis and reporting.
* Restructuring and Integration make it easier for the user to use for reporting and analysis.
* Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
* Data warehouse stores a large amount of historical data. This helps users to analyse different time periods and trends to make future predictions.

**A) Microsoft Excel Input to Microsoft Excel Output**

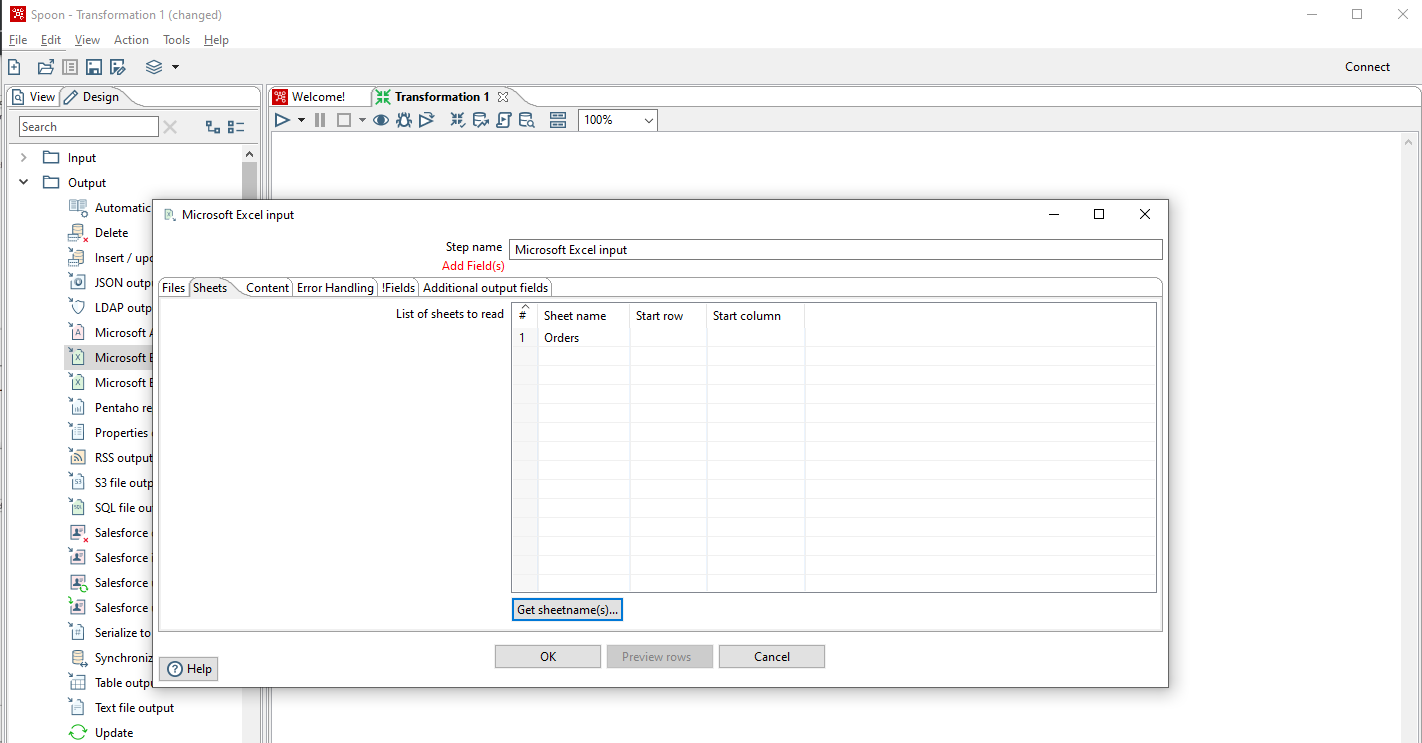
1) Drag and drop Microsoft Excel Input from input field and Microsoft Excel Output from output field. Create a hop between them.



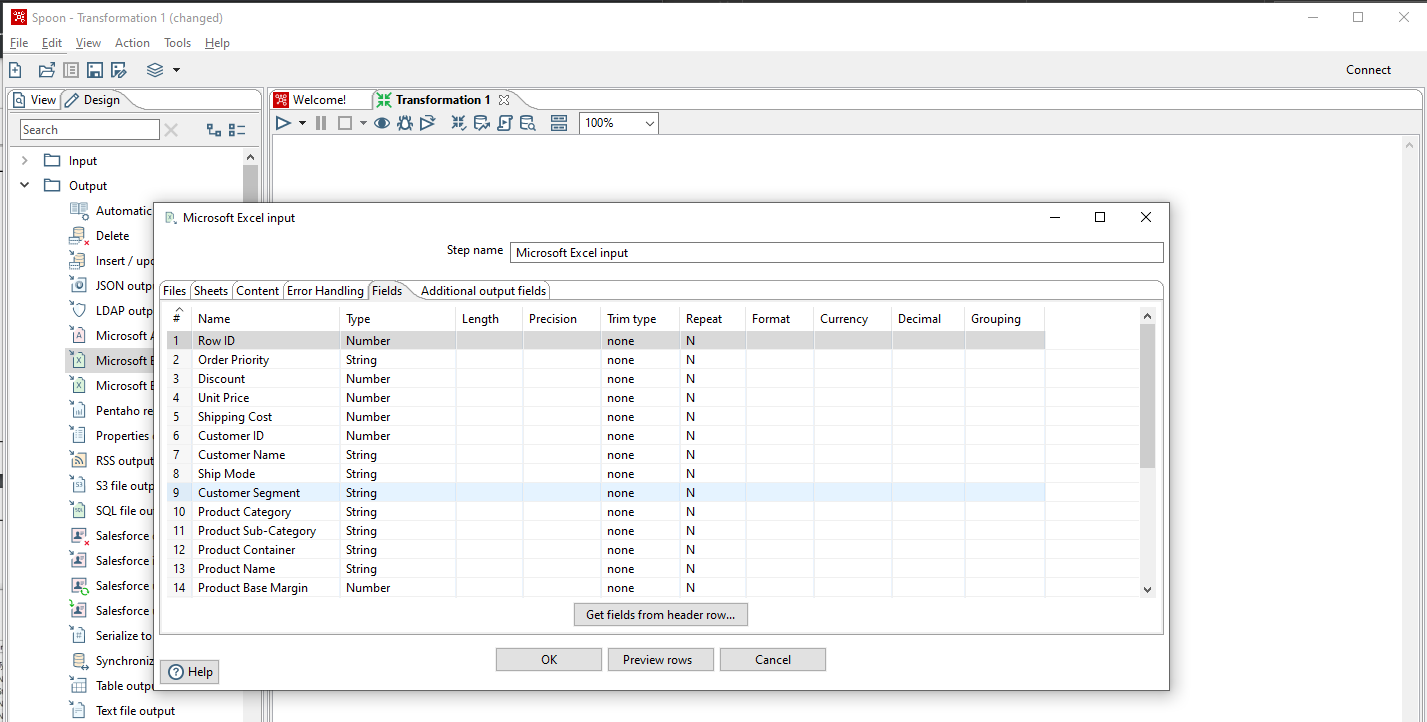
2) Double click on Microsoft Excel Input and browse the excel file and click on add.



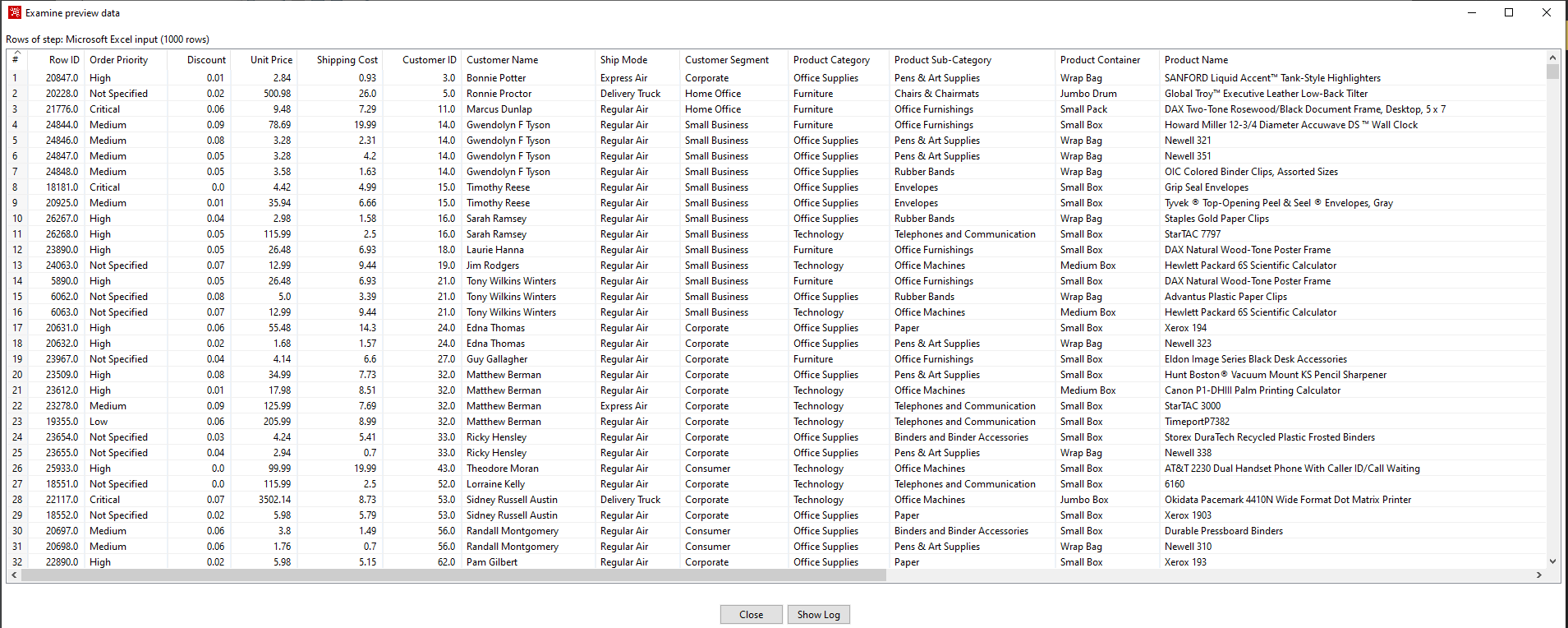
3) Click on Sheets -> Get sheet name. Select the Orders table.



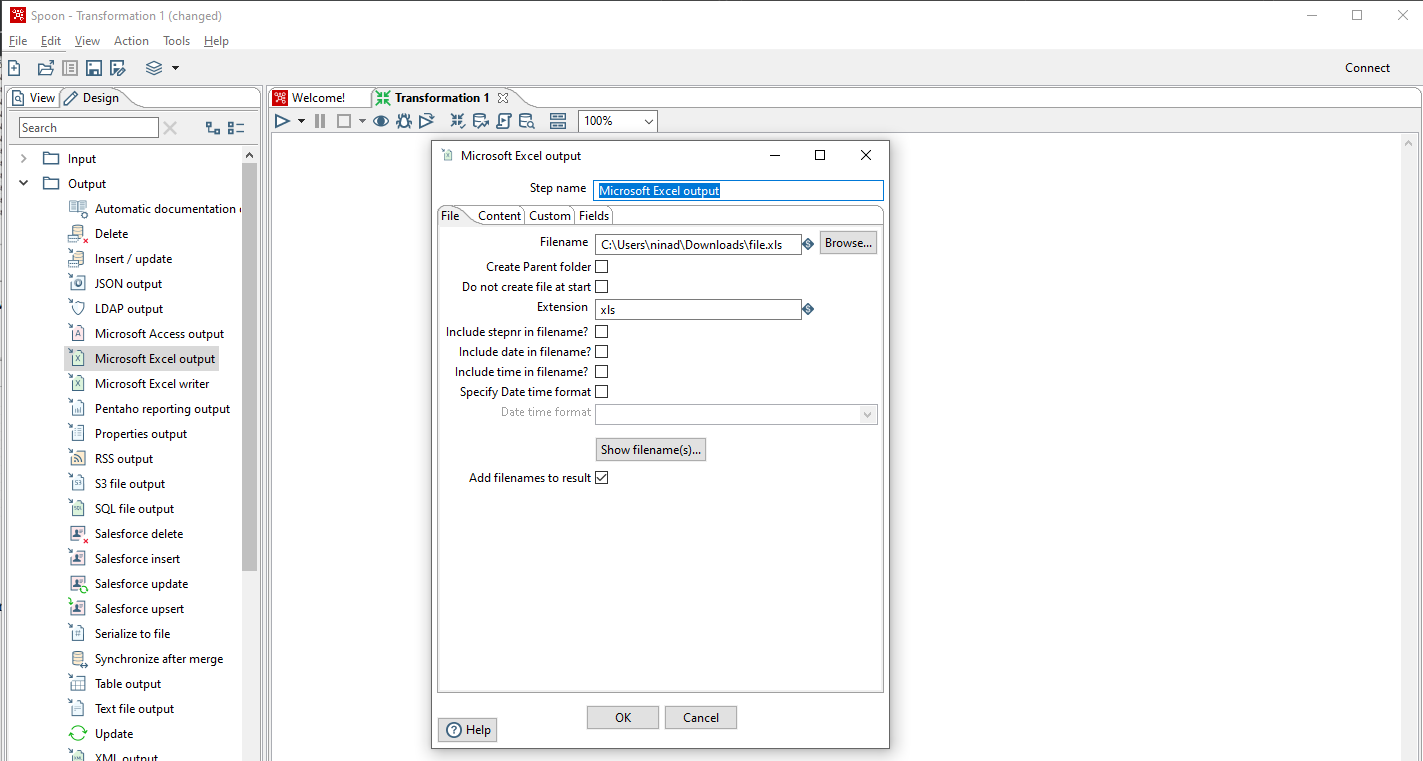
4) Click on Fields -> Get Field Name. The field names should appear as shown below.



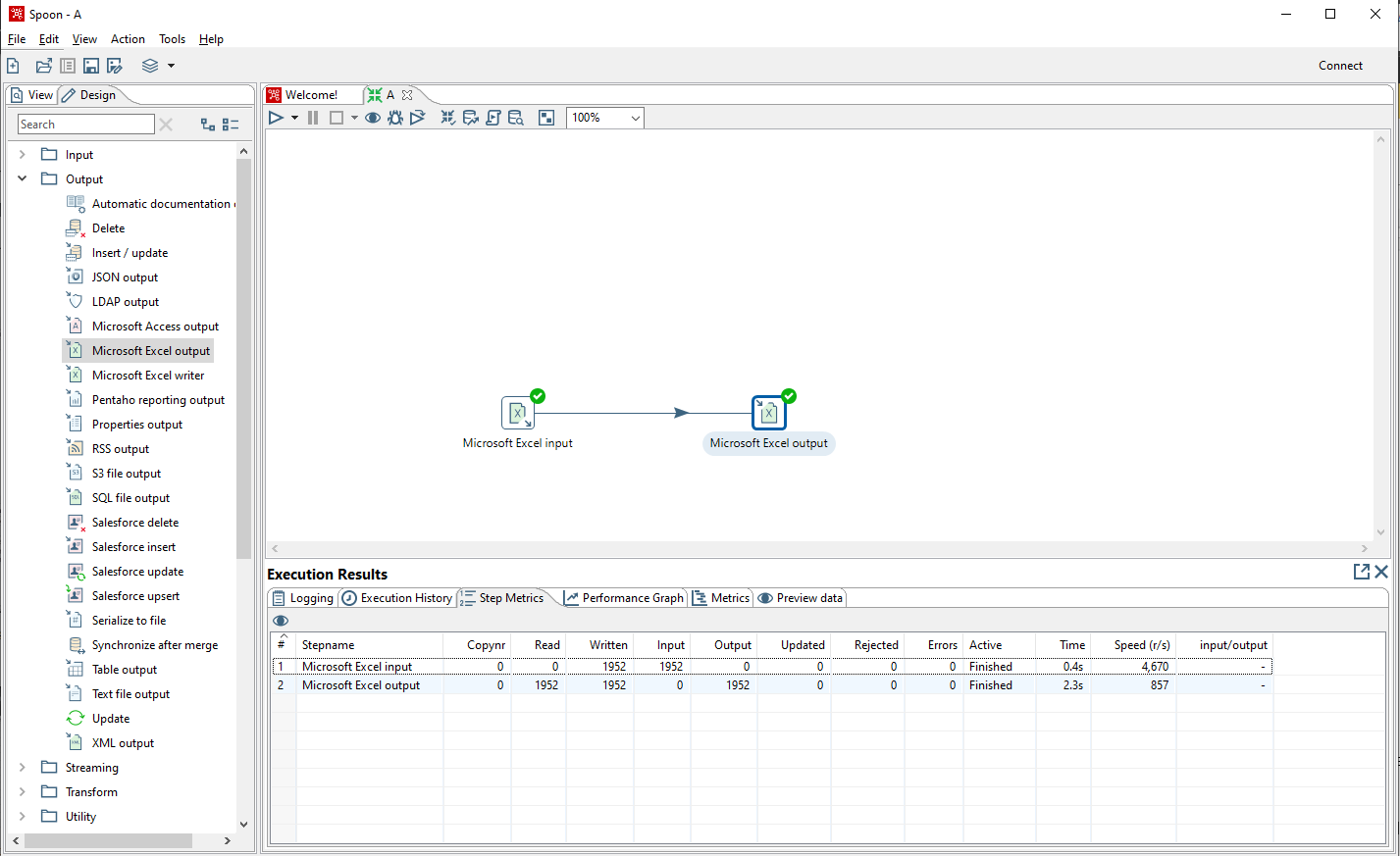
5) Preview the rows.



6) Double click on Microsoft Excel output. Click on Browse and save the output file.



7) Run the transformation.



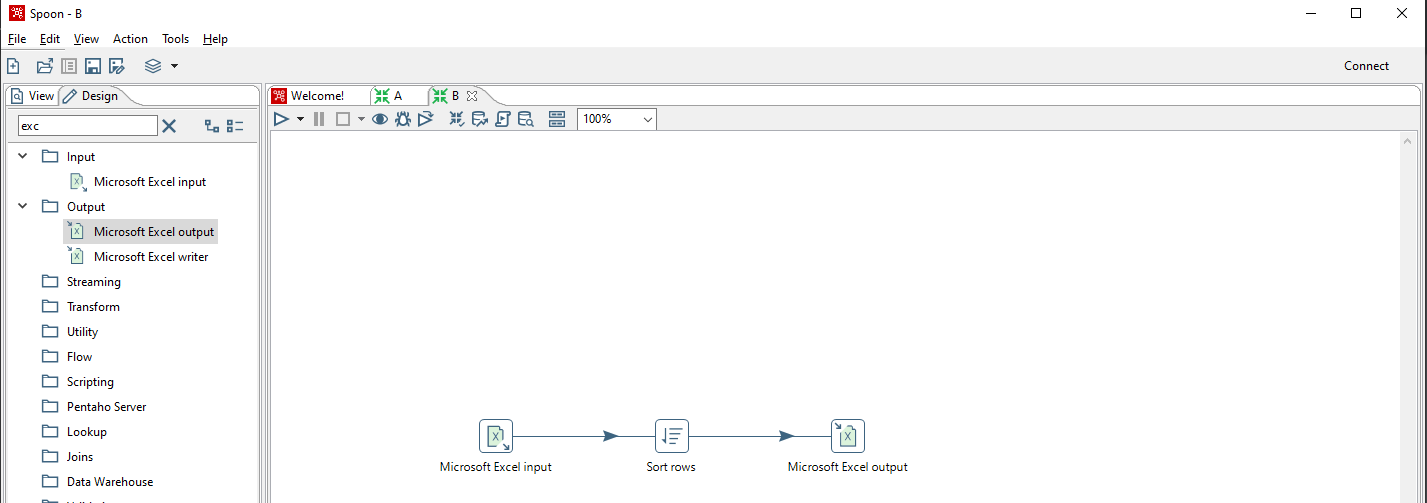
8) The output file should appear as shown below.



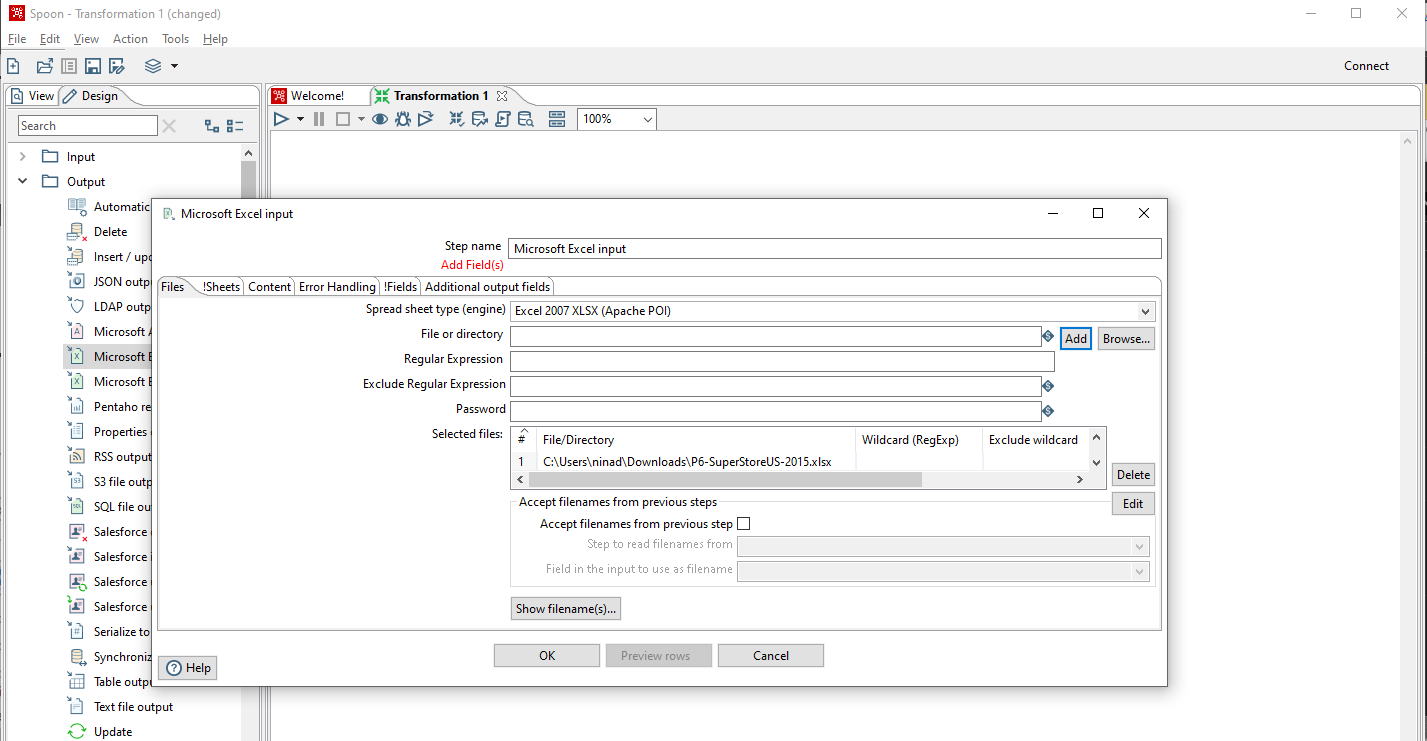
**B) Applying transformations to rows.**

**1) Sort**

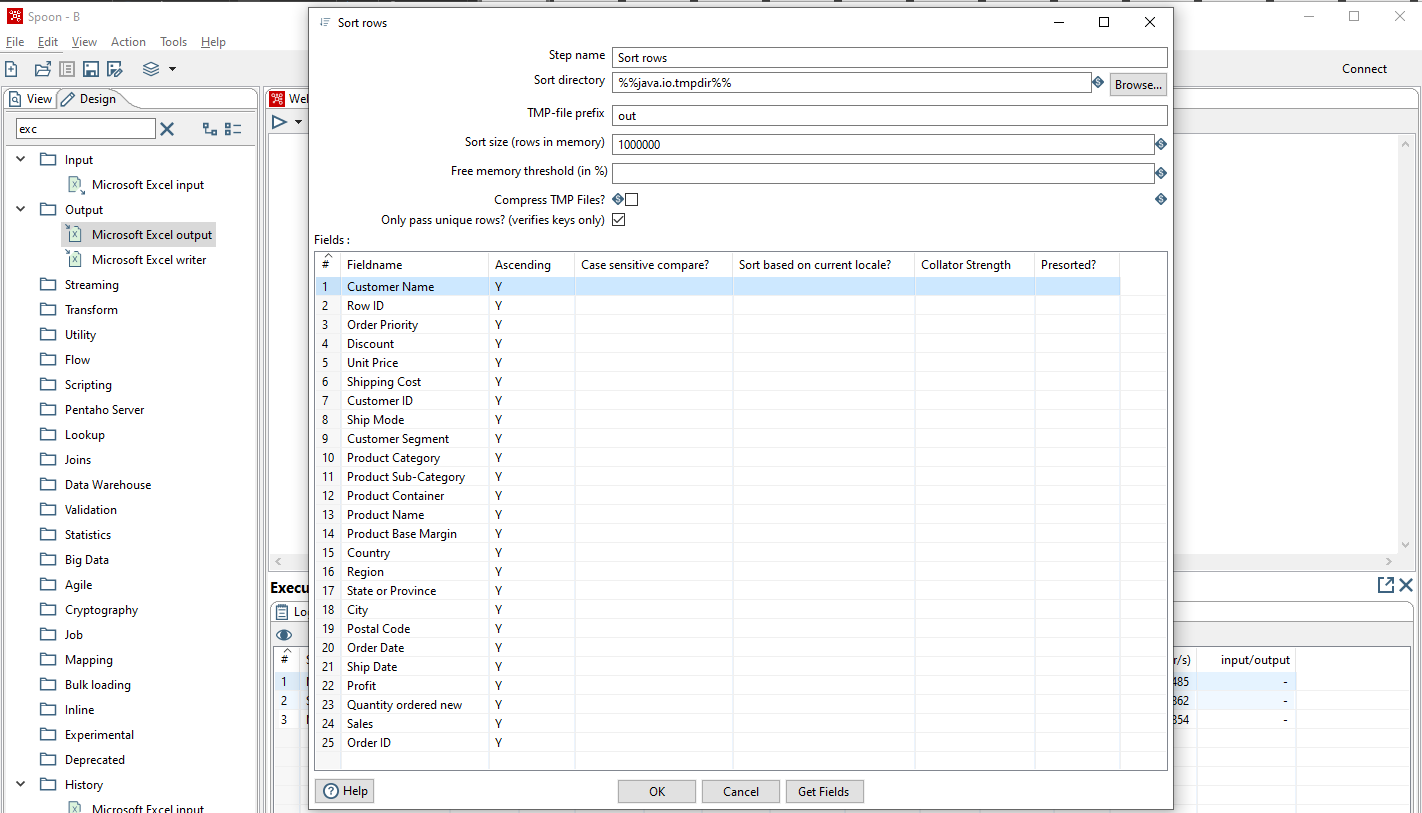
1) Drag and drop Microsoft Excel Input, Microsoft Excel Output and Sort from Transform field.



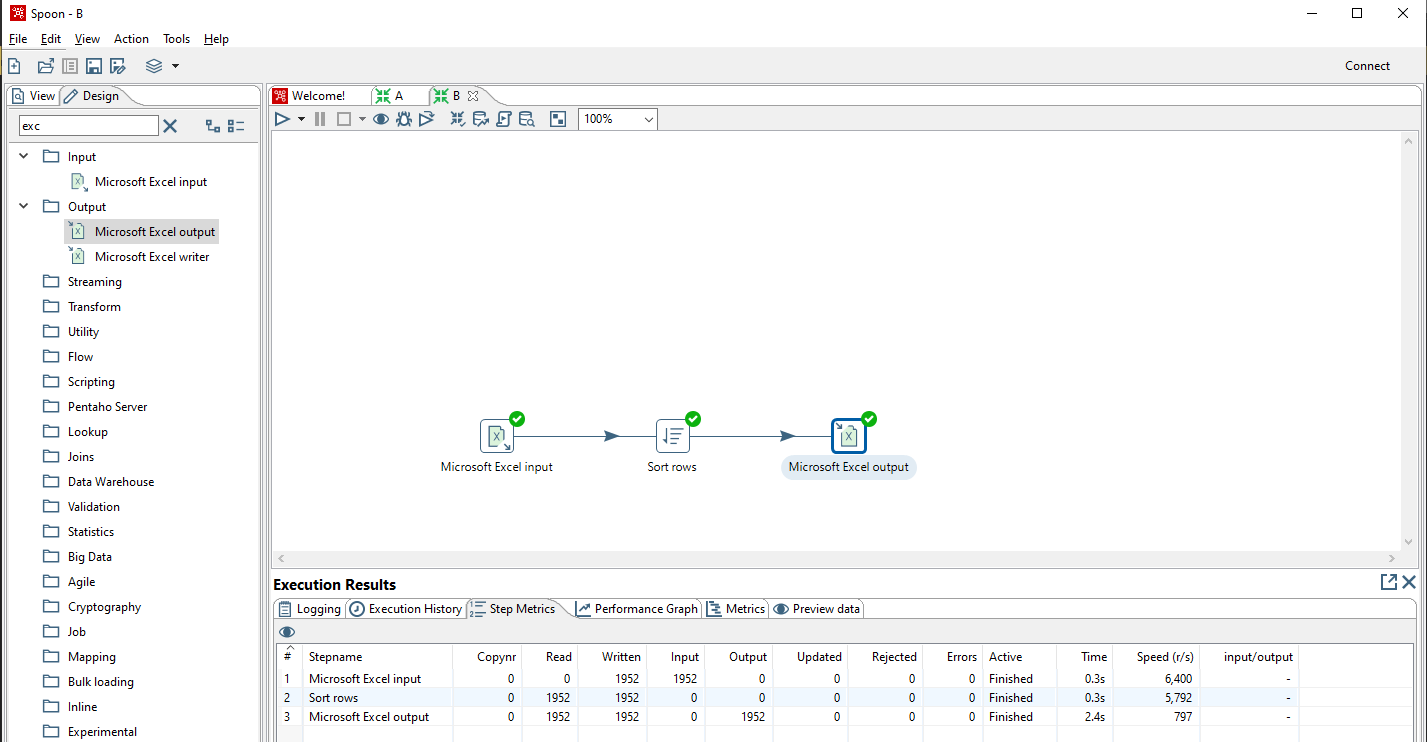
2) Set up the input file by choosing the appropriate Excel file. Choose Orders table and get field names. Preview rows.

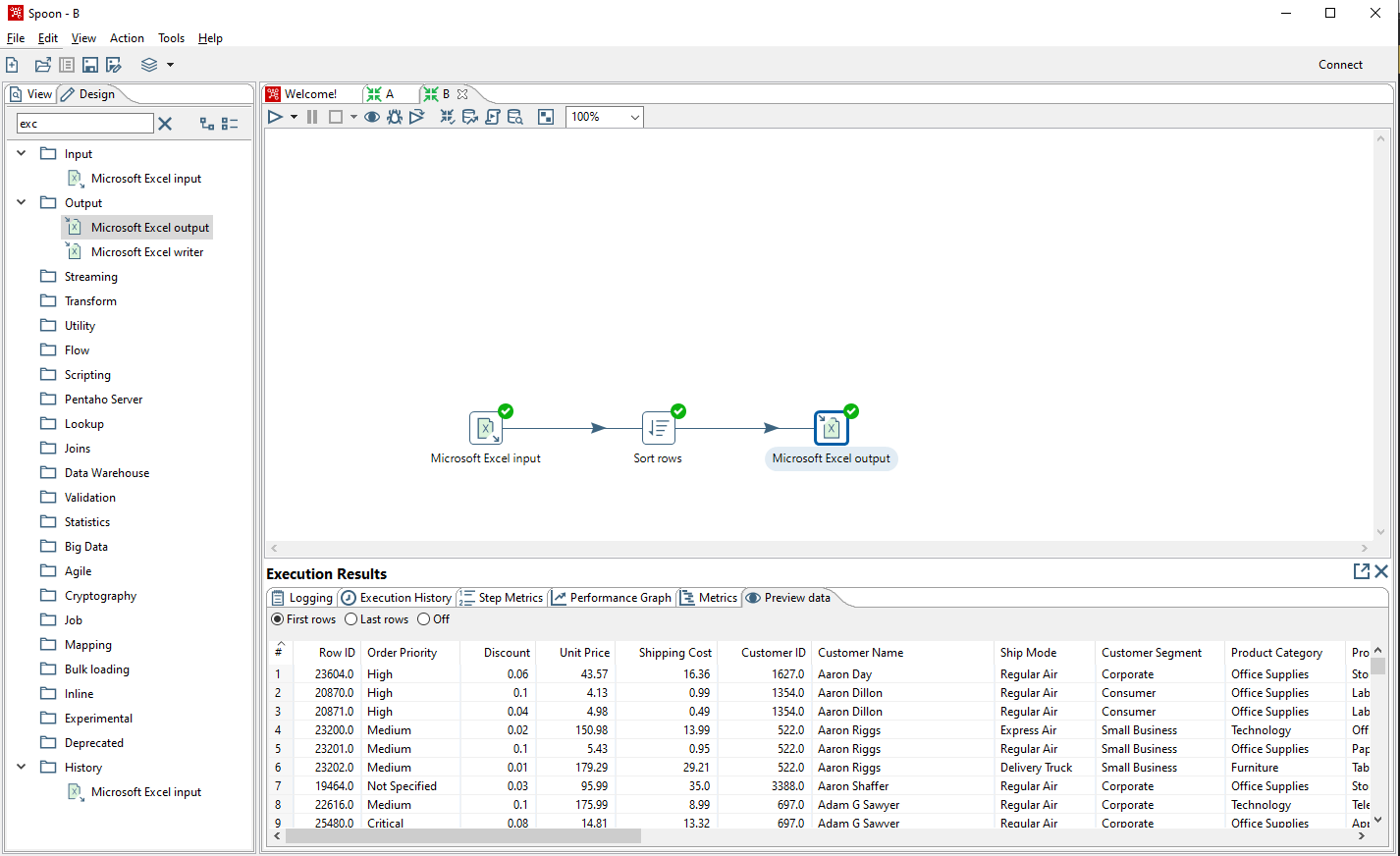


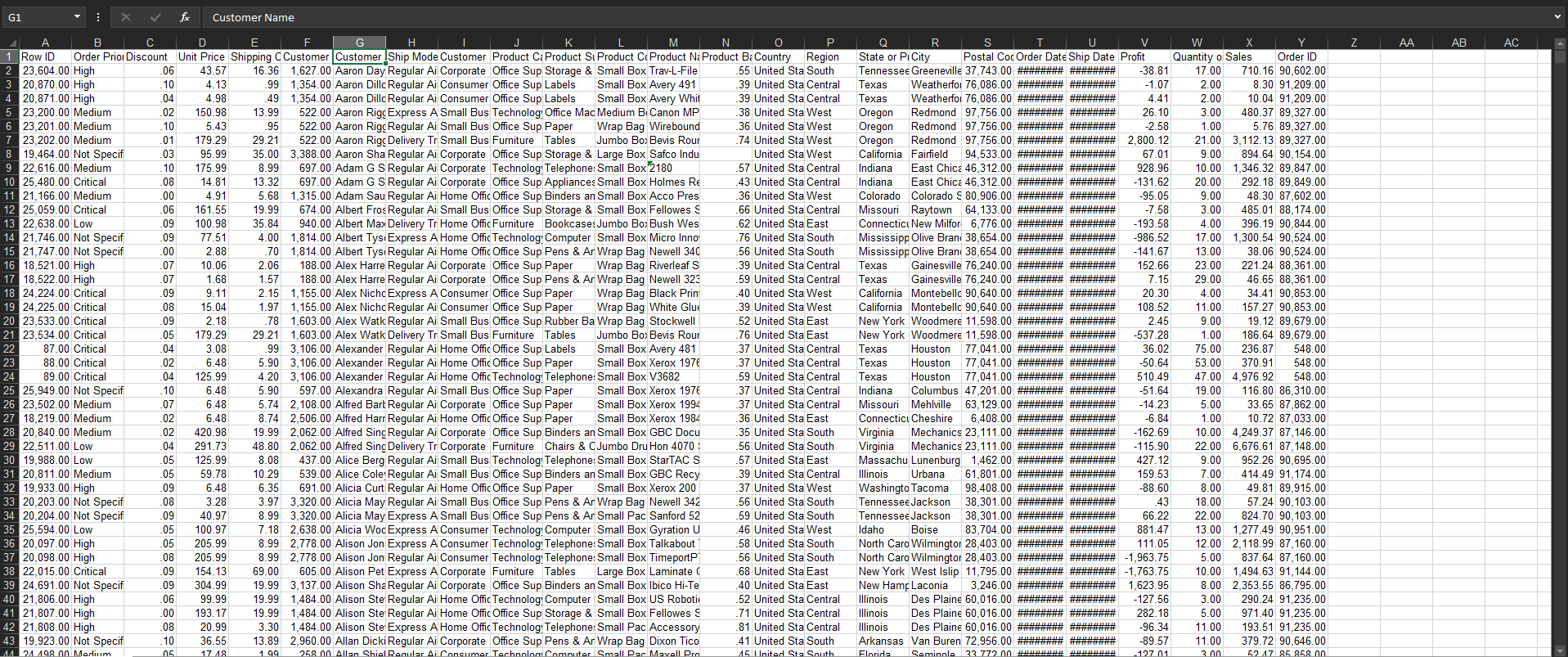
3) Double click on Sort -> Get Fields. Move up the Customer Names field.



4) Double click on the output file and save it. Run the transformation.

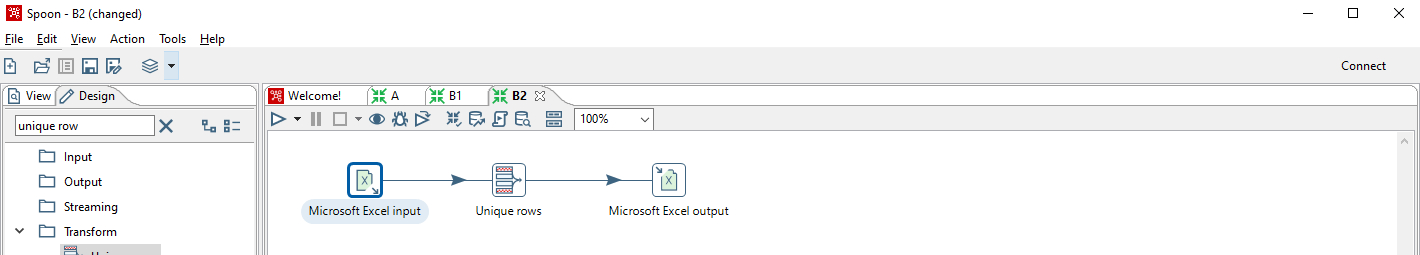




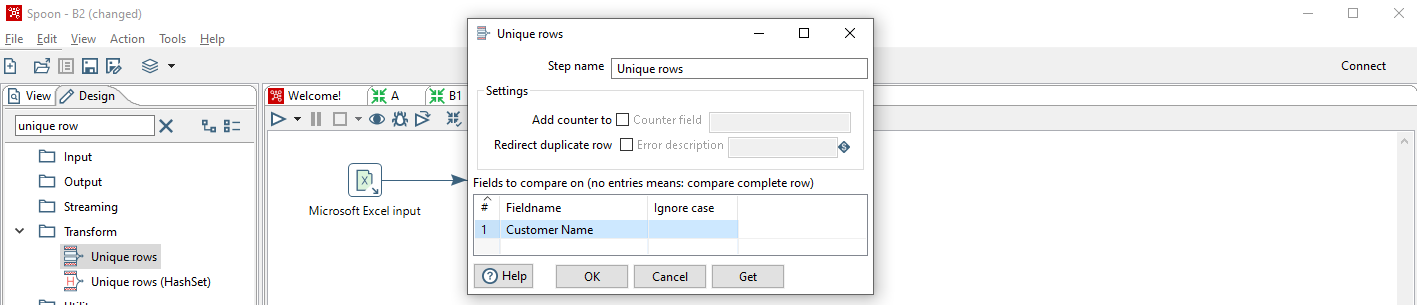


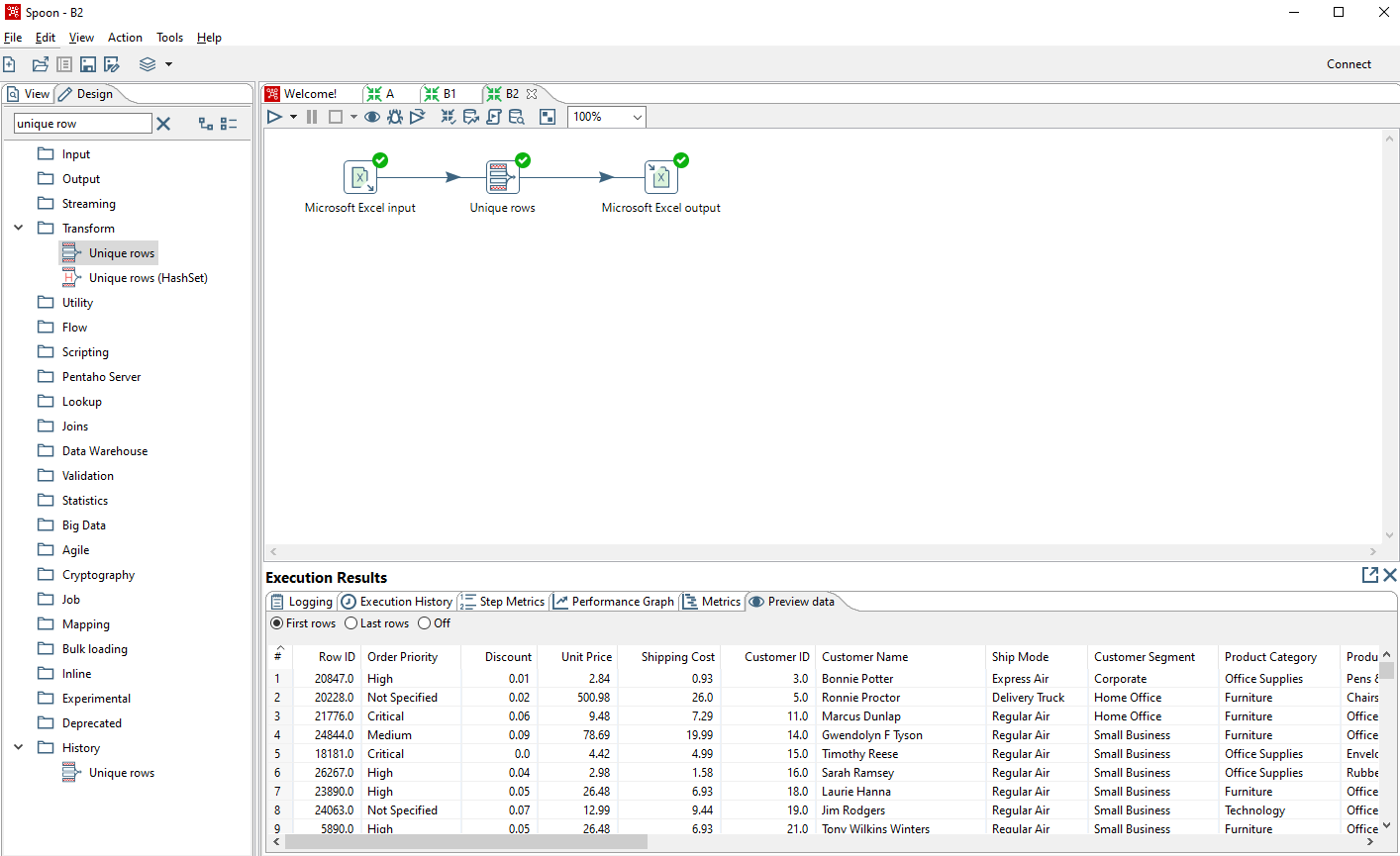
**2) Unique rows**

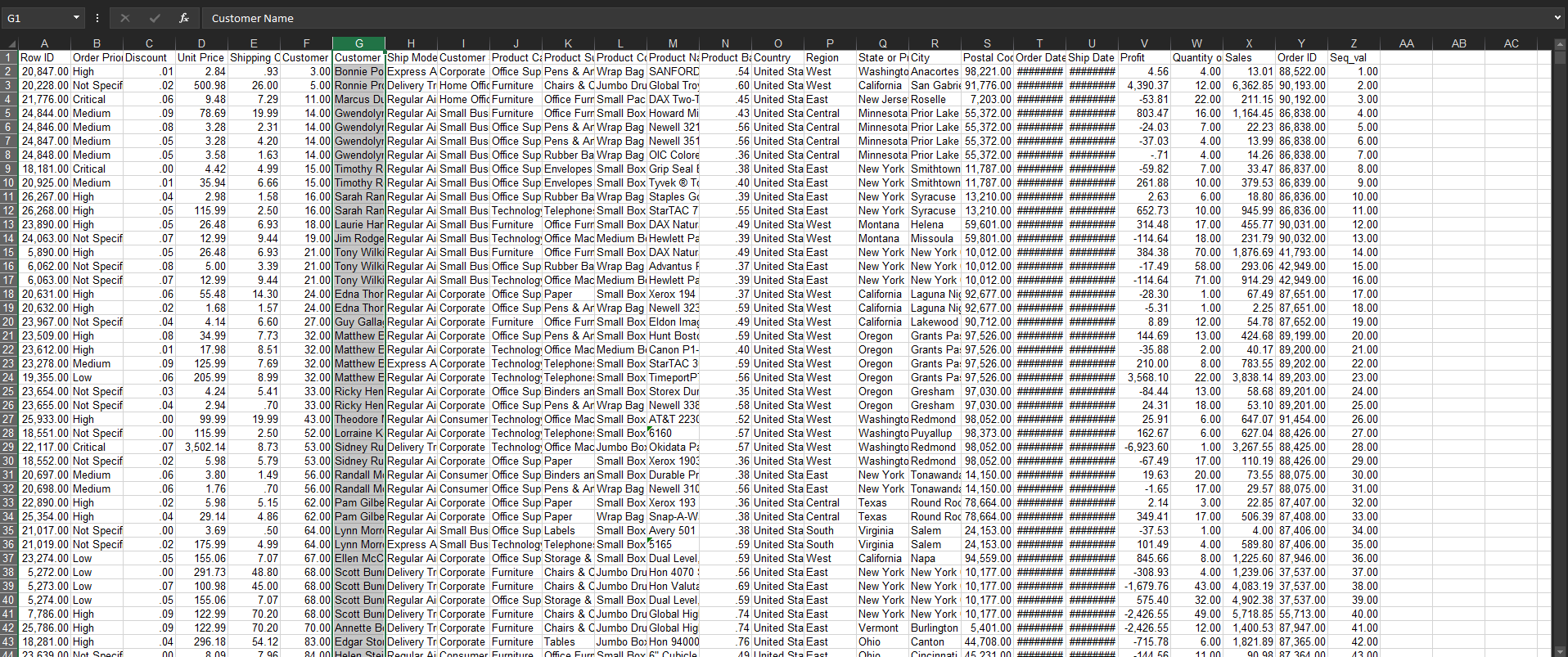
1) Drag and drop Microsoft Excel Input, Microsoft Excel Output and Unique rows from transform field.



2) In Unique rows, select the column for which unique values need to be displayed. Run the transformation.

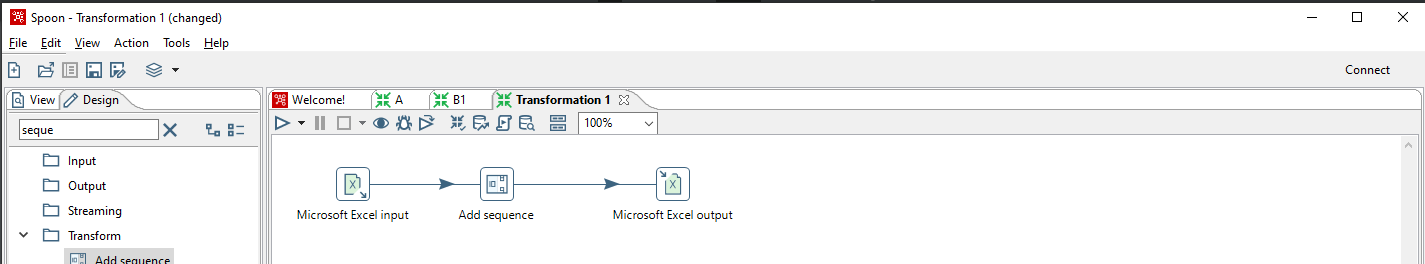




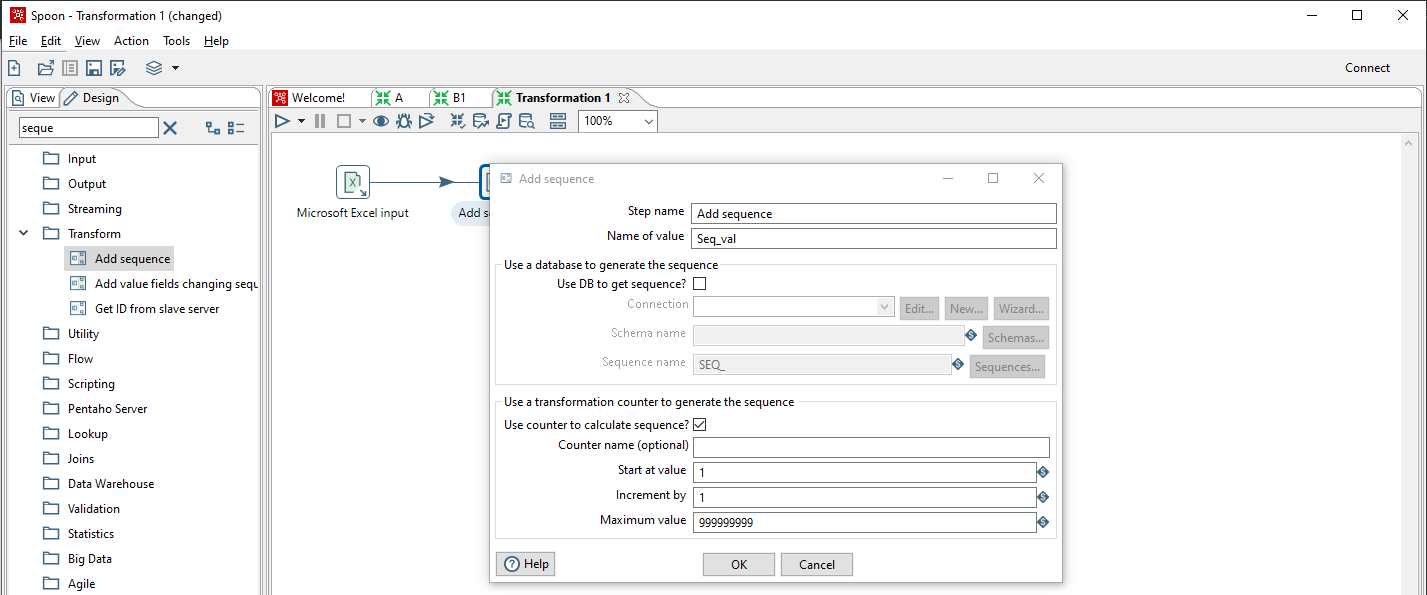


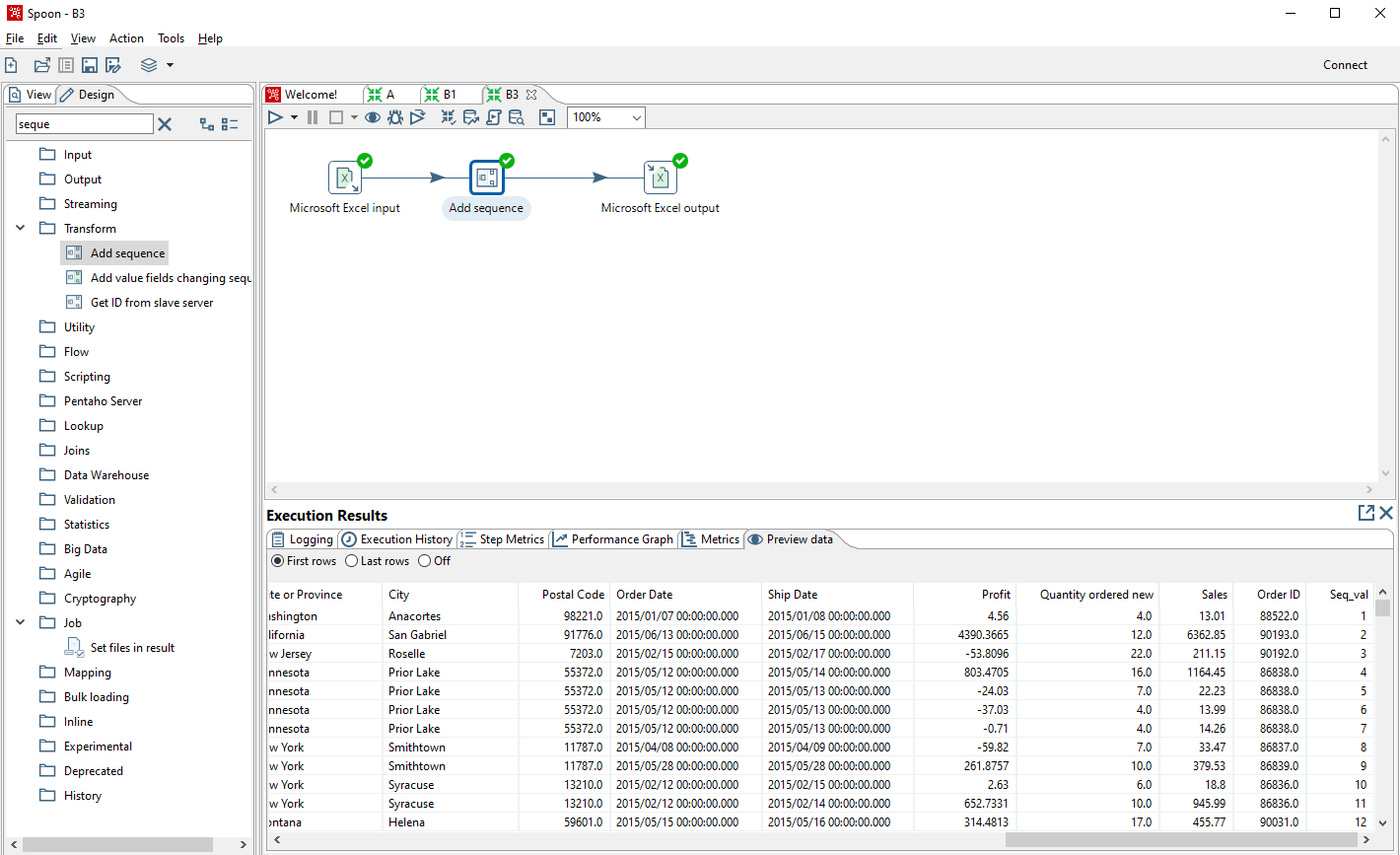
**3) Add sequence to the file**

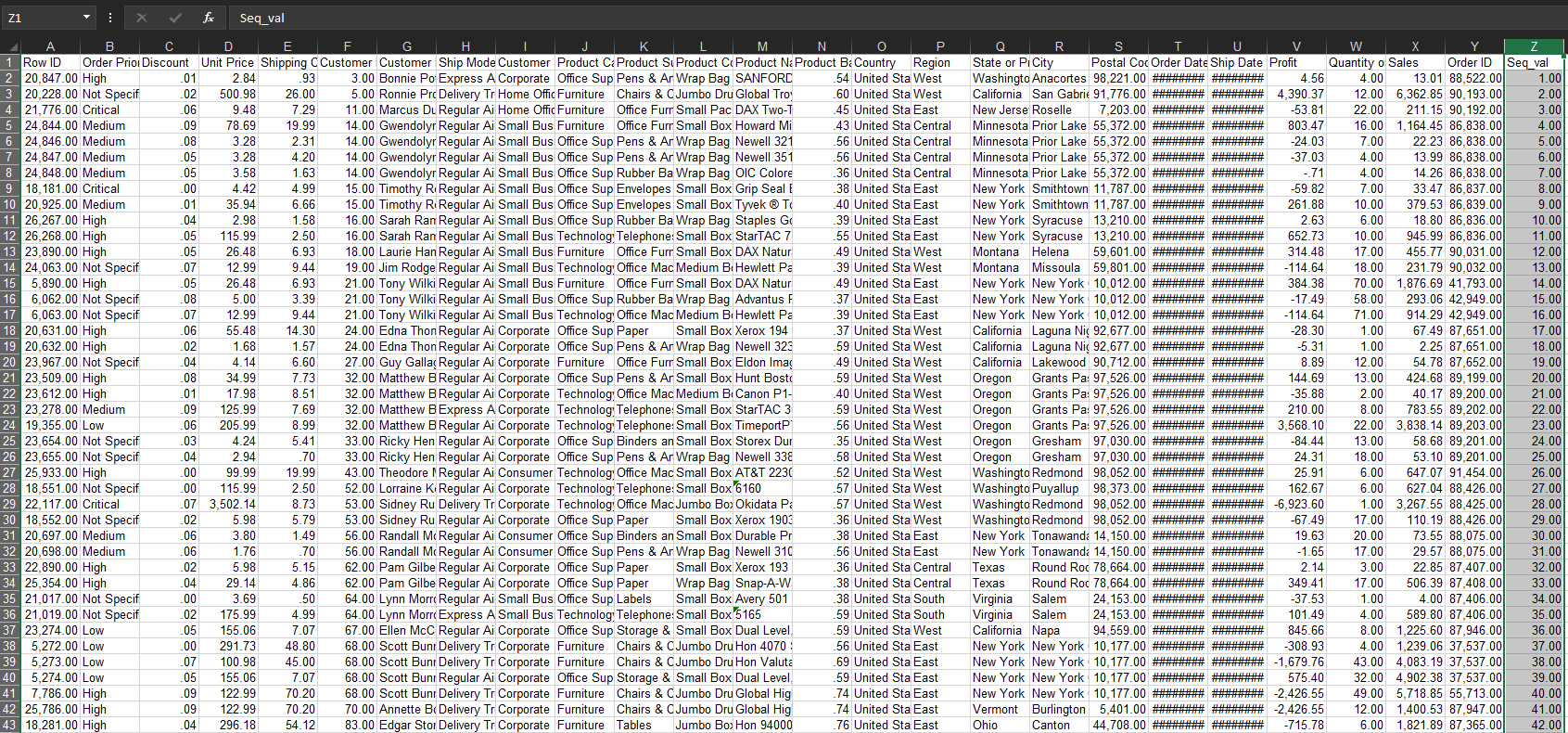
1) Drag and drop Microsoft Excel Input, Microsoft Excel Output and Unique rows from transform field.



2) In Add Sequence, select the start value and increment value. Run the transformation.

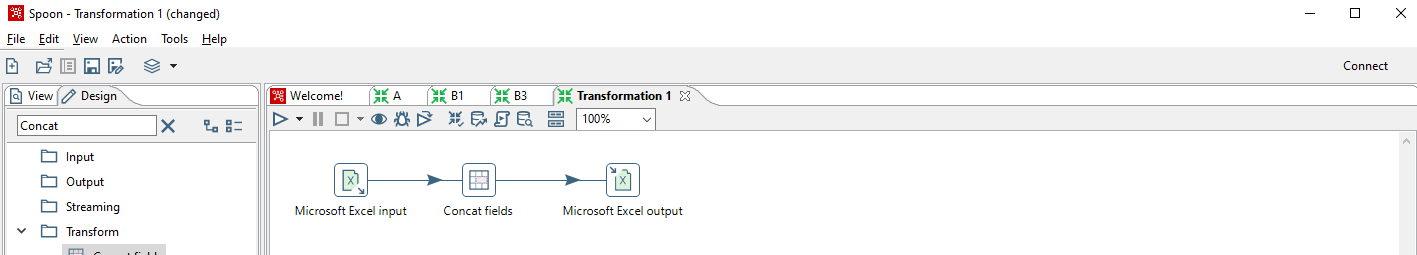




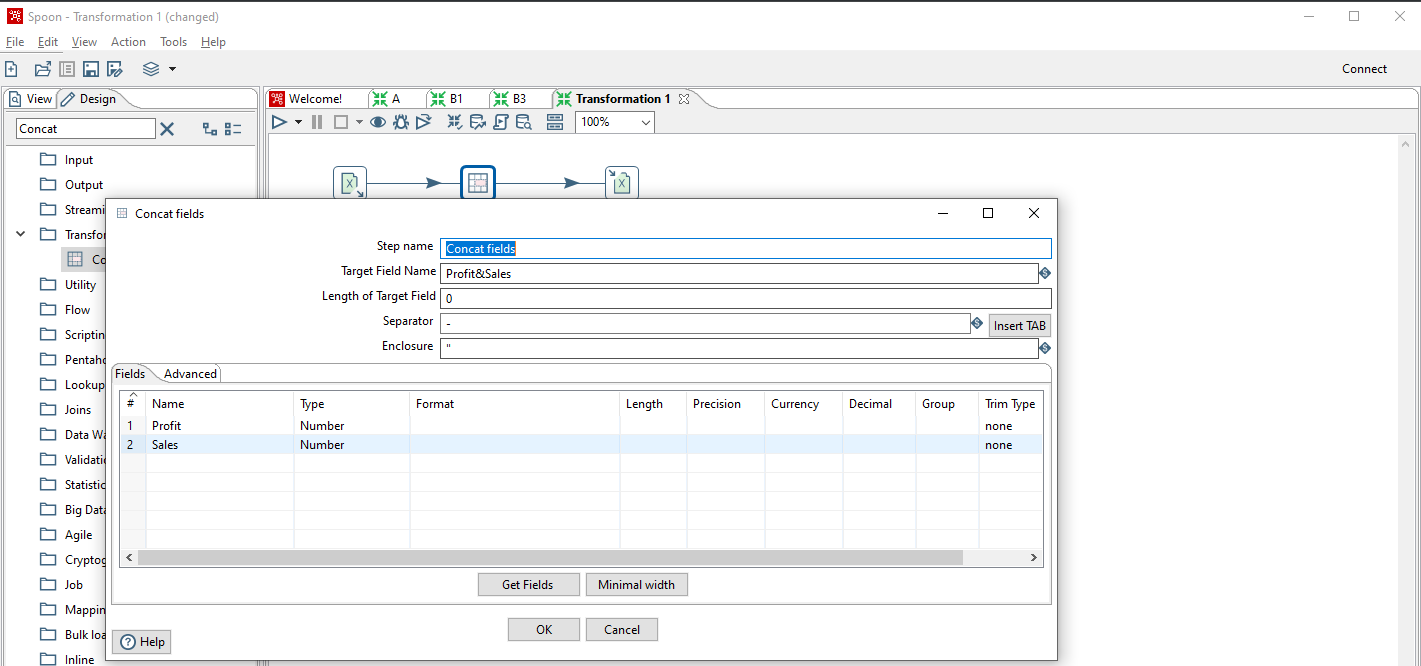


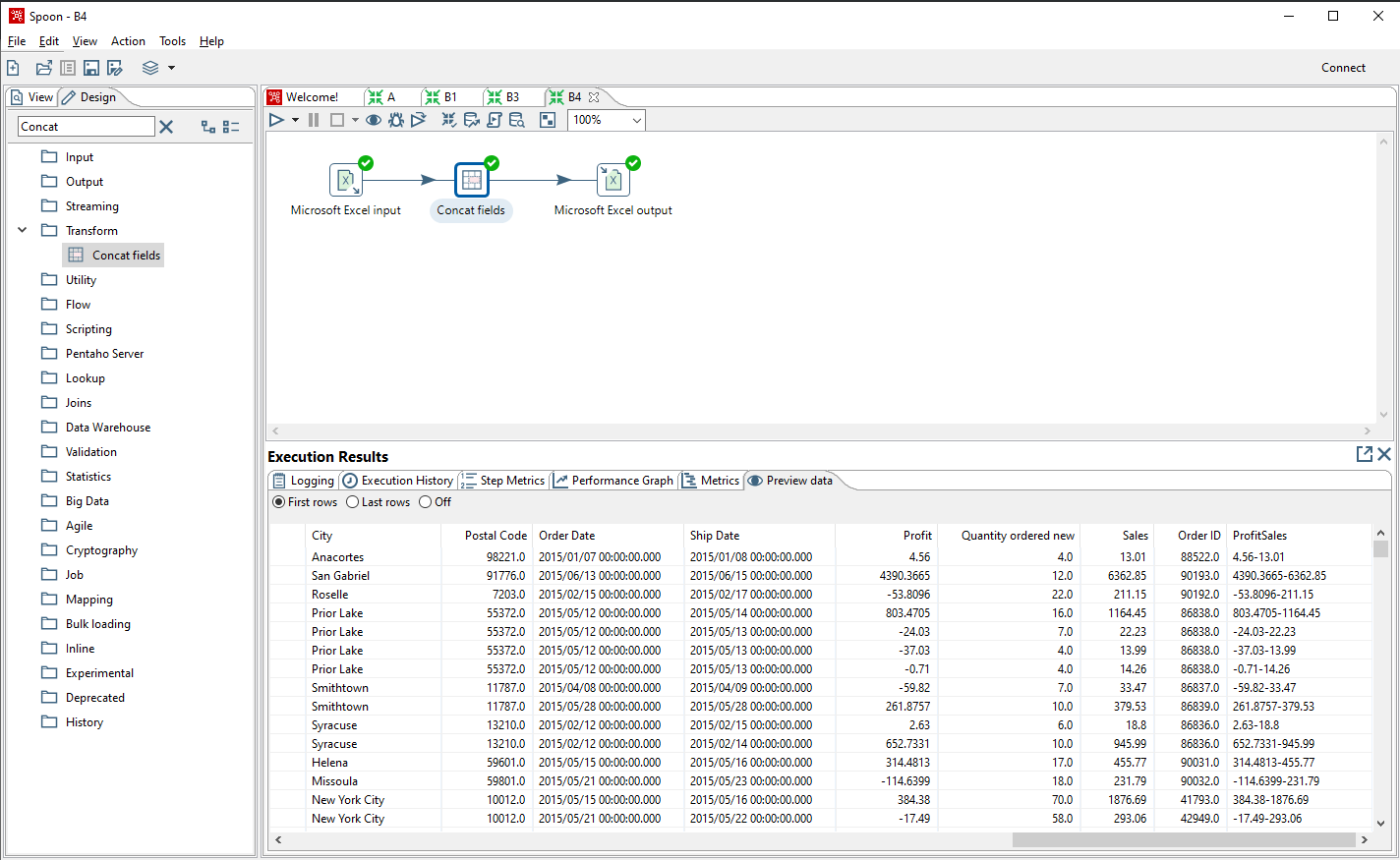
**4) Concatenate**

1) Drag and drop Microsoft Excel Input, Microsoft Excel Output and Concat fields from transform field.



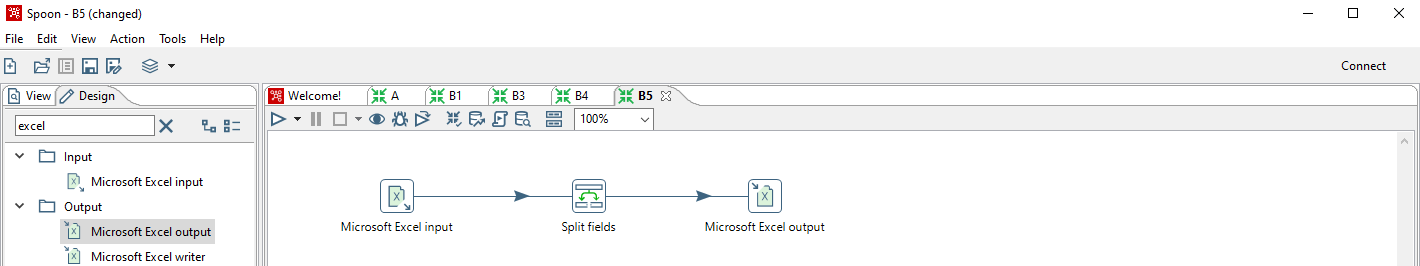
2) Select the fields to concat. Run transformation.



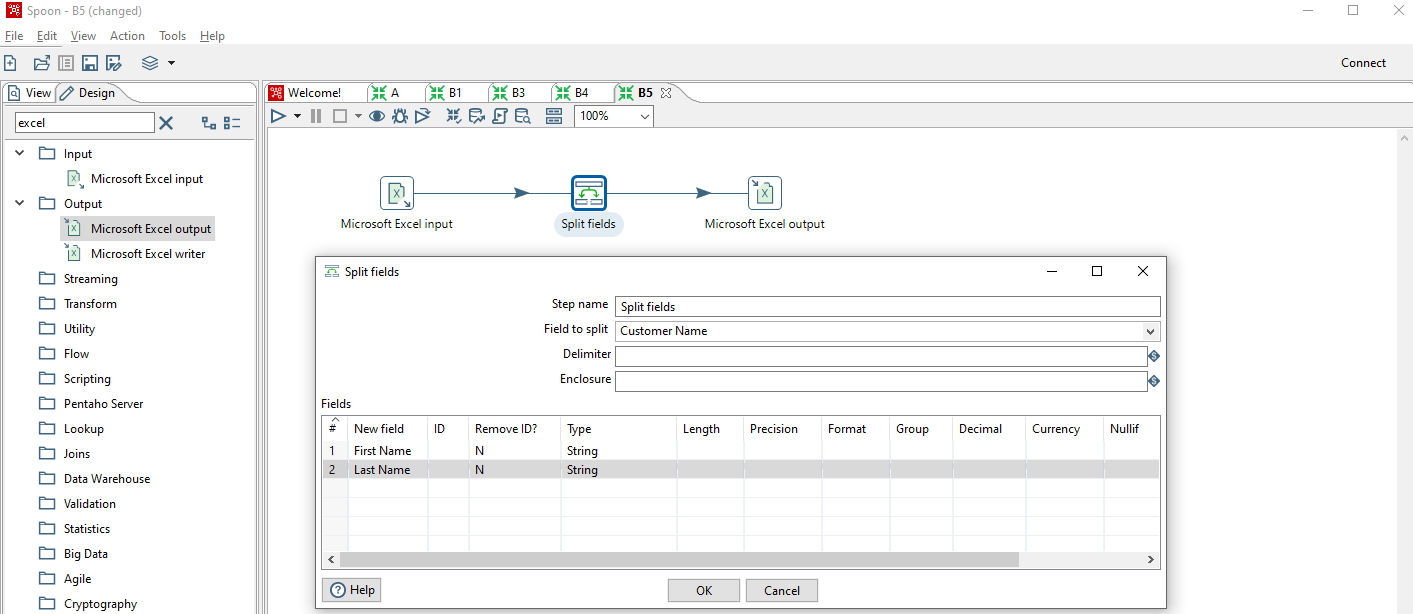


**5) Split**

1) Drag and drop Microsoft Excel Input, Microsoft Excel Output and Split fields from transform field.



2) In Split fields, choose the row you want to split and specify the delimiter. In this, the Customer name is split into first and last name using the “space” delimiter.



3) Run the transformation.

