– Welcome to DICE –
The Data Integration and Cleansing Environment based on ADOxx

Quick Guide to DICE 1.0

Christoph Moser

Vienna, June 10th, 2016

book.omilab.org

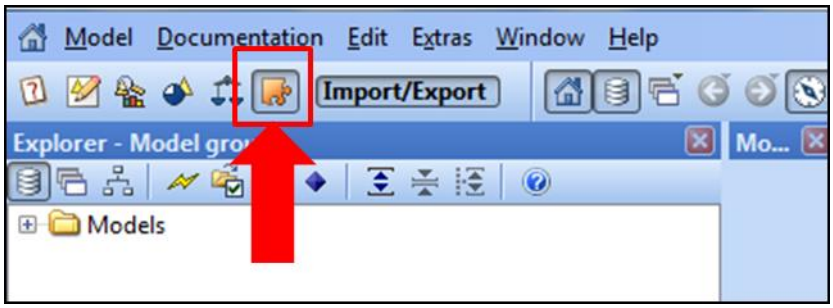
DICE 1.0 User Guide

After successful installation of DICE 1.0 you can import the DICE 1.0 demo data integration model. This model leads you through the most important feature of DICE.

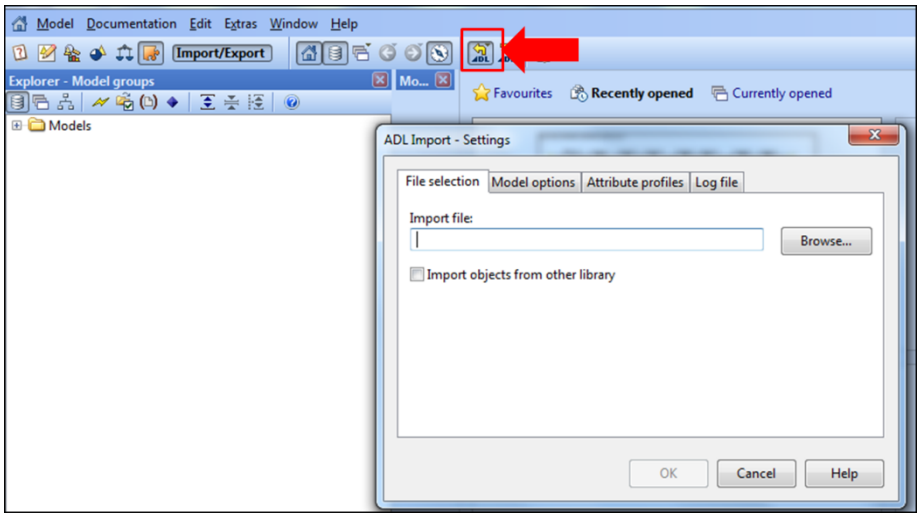
1 Perform your first data transformation task

Step 1: Import sample model

In order to import the sample model you have to switch to the DICE import/export component.

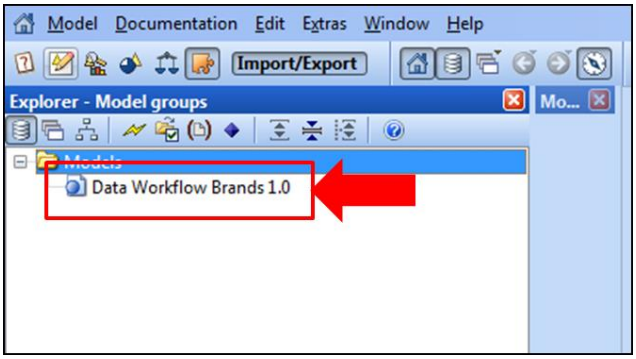


Chose the sample model “DICE sample model.adl” for import.

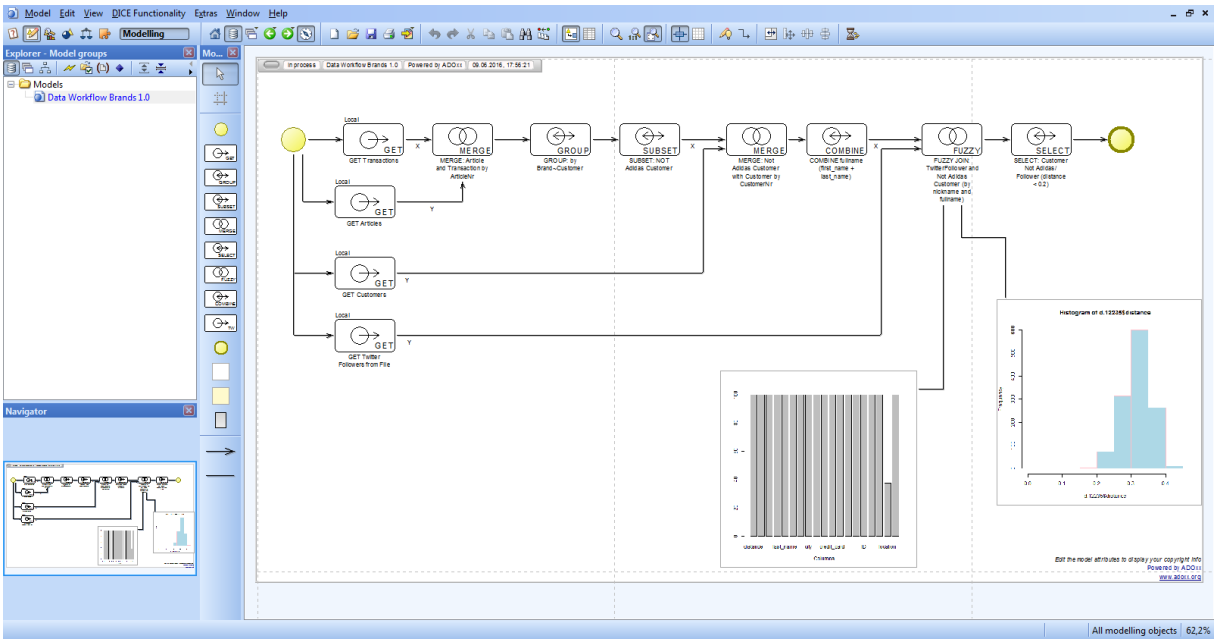


Step 3: Open the sample model

After successfully importing the sample model you can open it by double-clicking on the model within the explorer.



The model is opened in the DICE modelling environment

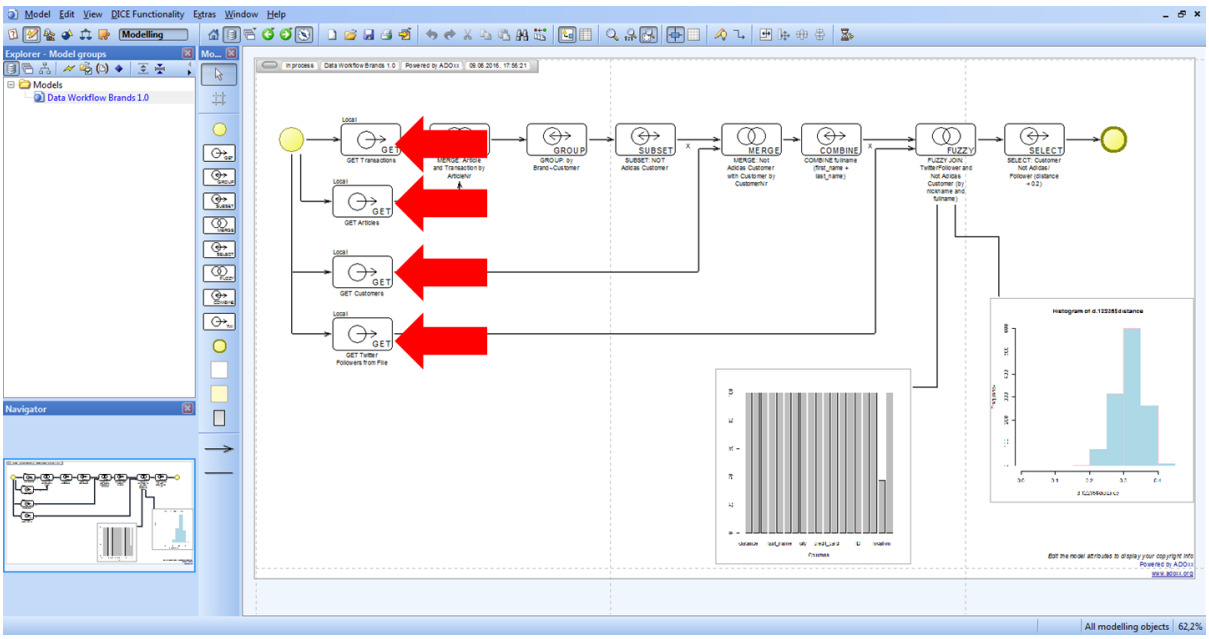


Step 5: Assign sample input data files

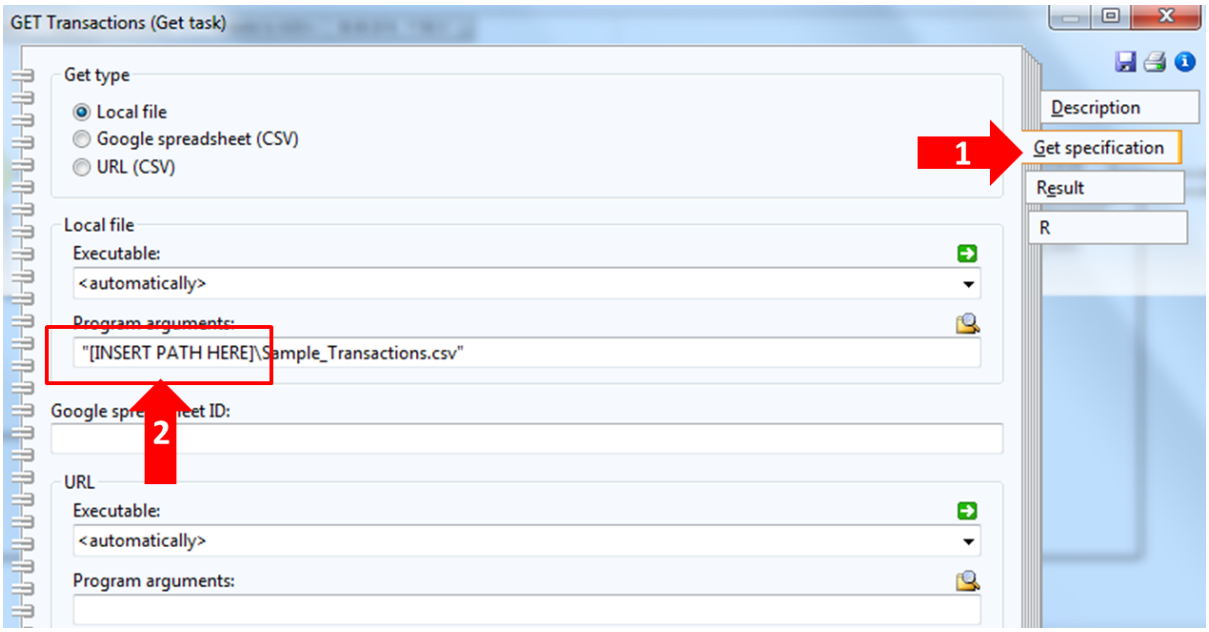
Assign input data files for all GET-transformation tasks:

- GET Transactions
- GET Articles
- GET Customers
- GET Twitter Followers from File

You can open the notebooks (property boxes) of these transformation tasks by double-clicking on the shapes.

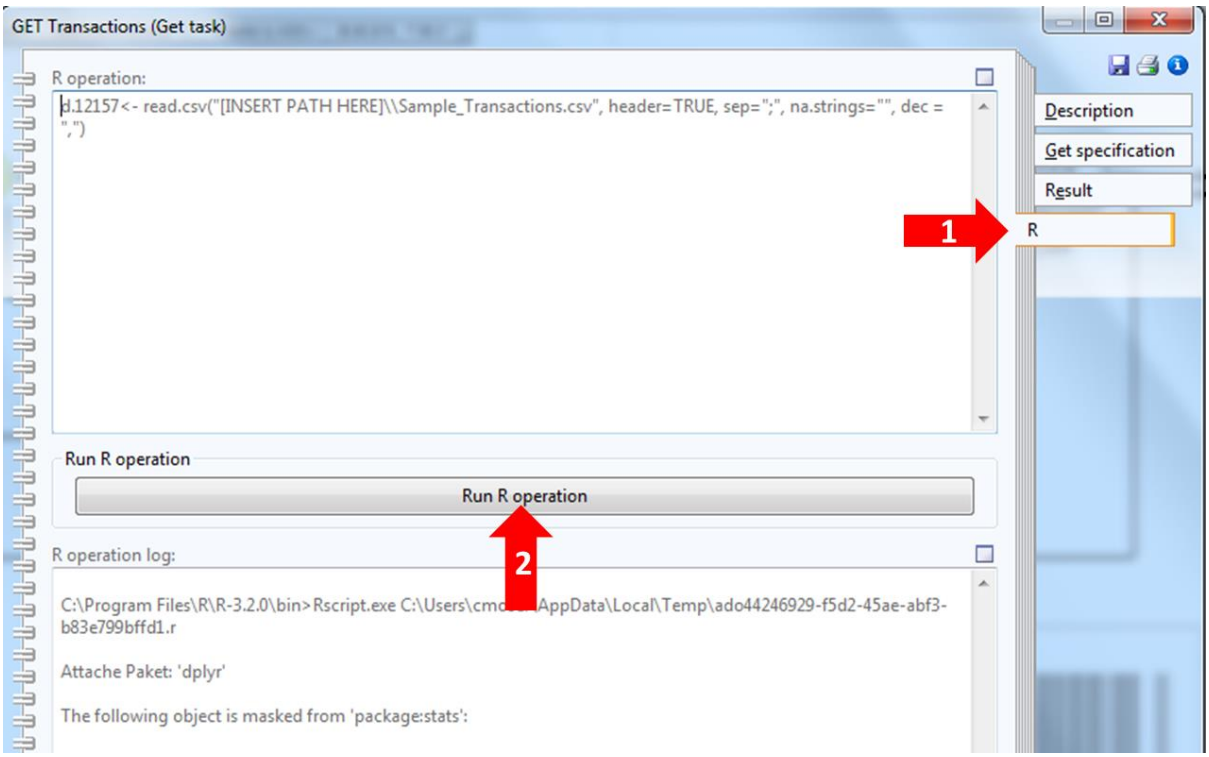


Switch to the chapter “Get Specification” and replace the path to the sample input data files.

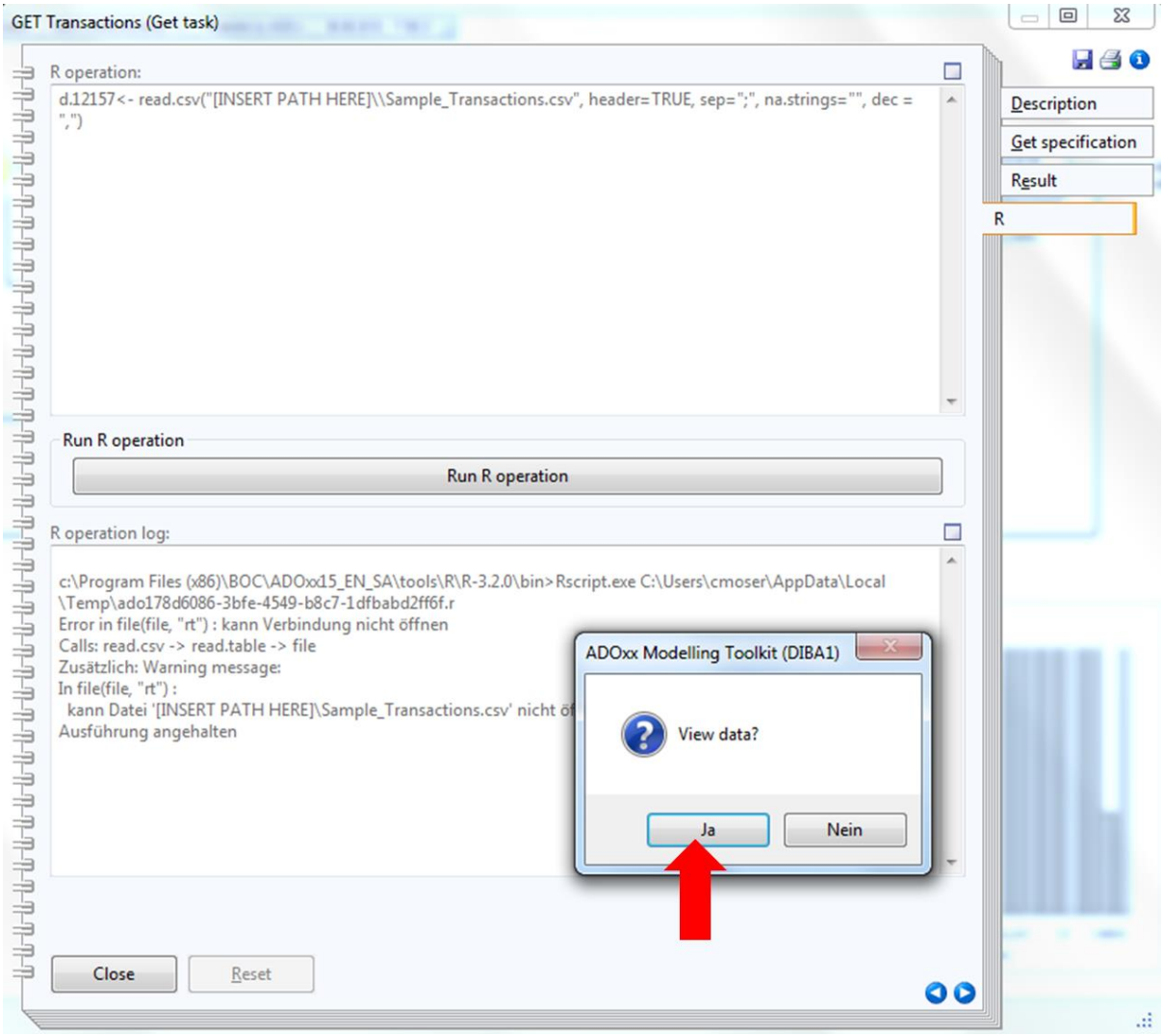


Step 6 – Run your first transformation task

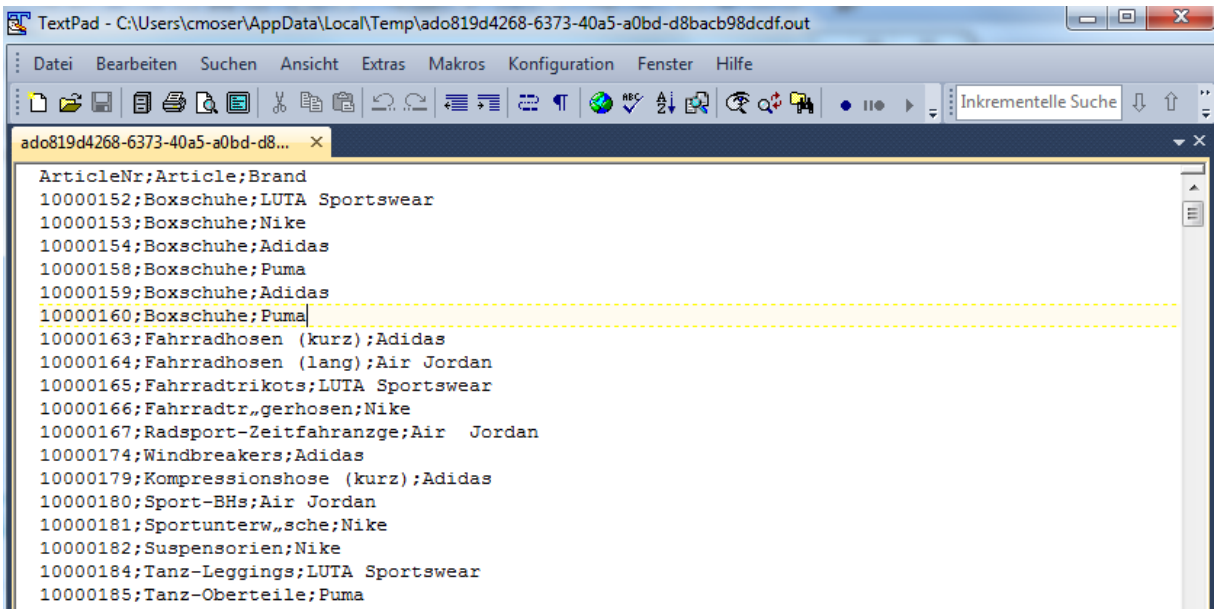
By switching to chapter “R” and clicking on the button “Run R operation” the input data is loaded for further processing.



You can view the data by clicking on “View data” after the sample data has been read.

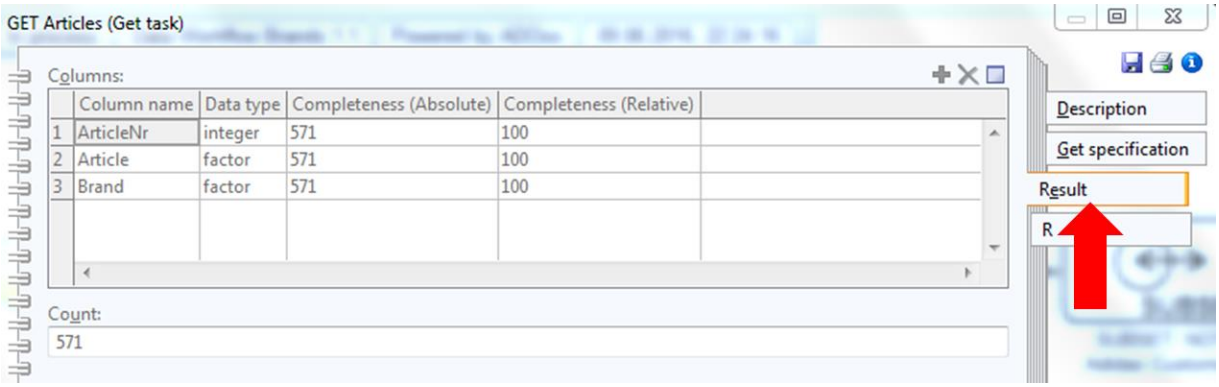


The resulting data frame is opened in your preferred text editor.



Step 7 – Check metadata

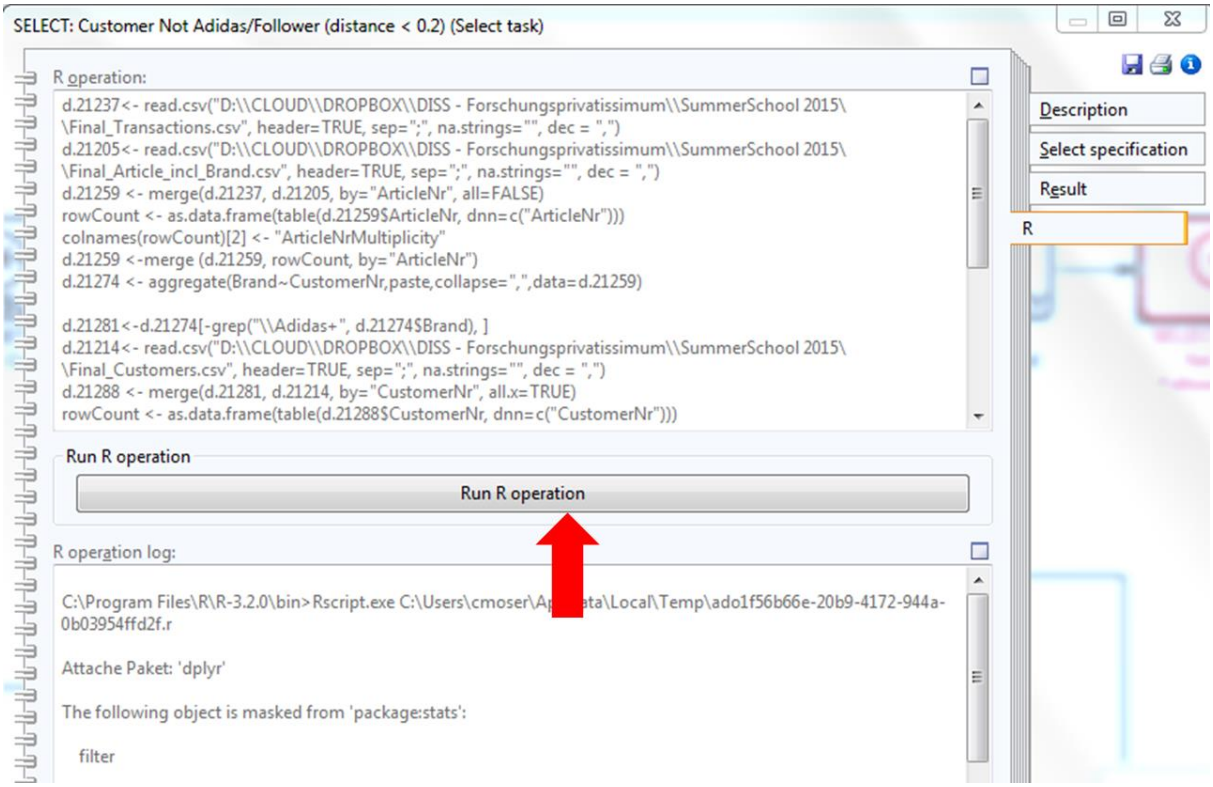
You can access the metadata generated by performing this simple load task by switching to chapter "Result". The table shows column names, data types, as well as absolute and relative completeness of the records within the chosen input file.



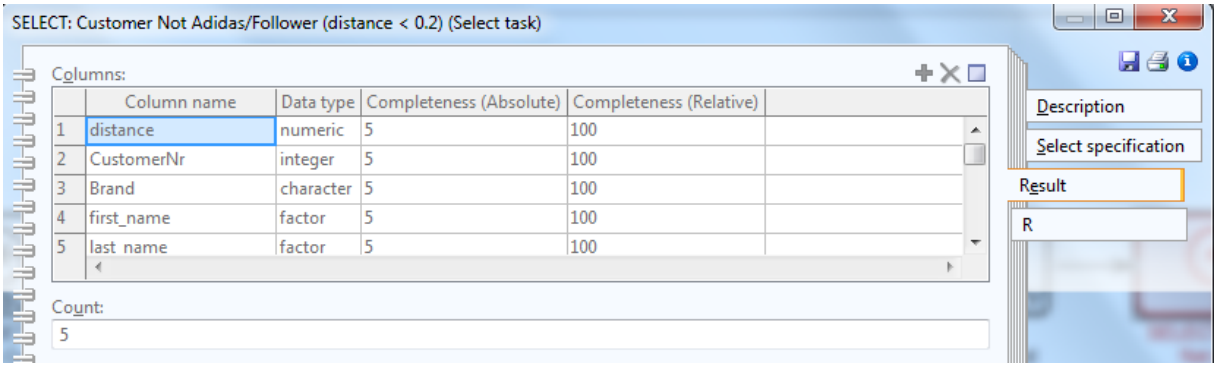
2 Run the entire data integration process

In order to run through the entire process and to perform all the defined transformation tasks you simply open the notebook of the final task "SELECT: Customer Not Adidas/Follower (distance < 0.2)" and switch to the chapter "R".

The attribute "R operation" shows the entire platform-specific code (the DICE implementation uses R statistics for the required calculations) which will be run. This code is automatically generated based on all proceeding transformation tasks.



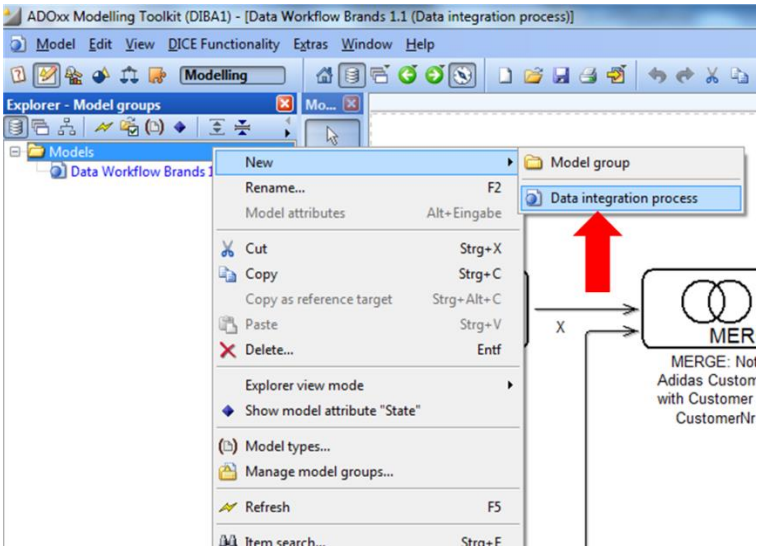
Switch to chapter “Result” to view the metadata of the resulting dataframe.



3 Create your own data integration process

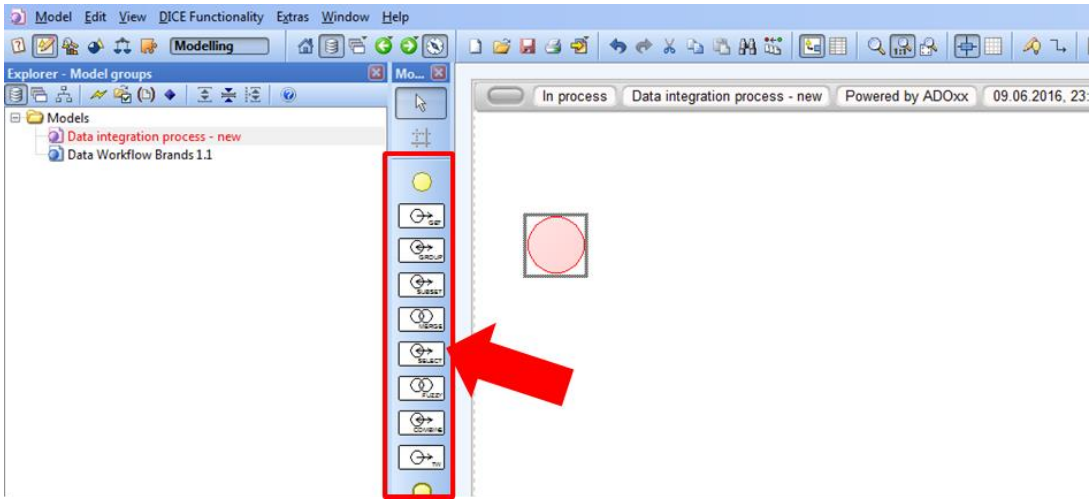
Step 1 – Create a blank model

Create a blank model of modeltype “Data integration process”



Step 2 – Model the data integration process

Insert a process start object first. Then create the GET transformation steps and connect them to the process start. Then create any further transformation steps.



For each of the transformation steps you have to define transformation tasks parameters in chapter "Get specification" of the respective transformation tasks.

The following transformation tasks are available in DICE 1.0:

- GET: Use this task for loading the input data files. Access data from your local drive, from google drive.
- GET TW: This is a special GET task, which retrieves and restructures twitter followers from www.twitter.com
- GROUP: Use this task to perform groupby-transformations on the given dataset.
- SUBSET: Use this task to perform select operations on the dataset.
- MERGE: This transformation task is used to merge two datasets. Be aware, that you have to define left and right dataframe in the incoming relations objects.

- FUZZY MERGE: This transformation task allows for fuzzy merge based on the Jaro-Winkler technique.
- COMBINE: Use this transformation task to combine two columns of a given dataset.

NOTE:
This first version of DICE provides some of the most important basic transformation task types. It is planned to enrich the set of transformation tasks in the upcoming versions. If you are interested: JOIN OUR TEAM.