Microsoft

Data Science Essentials

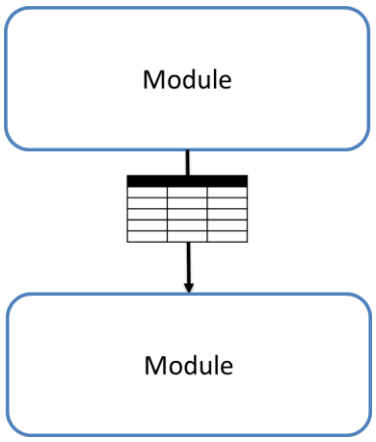
Data Exploration

A large proportion of a data scientist's work is to explore data and metadata. Usually this type of work is performed by writing code to manipulate data in R or Python, or by using some of the built-in modules in Azure Machine Learning.

Most data exploration involves loading the data into tabular structures, much like tables in a relational database or in a spreadsheet, and then performing operations on these data tables to view the data they contain. Languages like R and Python use a data structure called a *data frame* to encapsulate a table of data, and each language provides a set of functions that you can use to work with data in data frames.

Tables in Azure Machine Learning

Azure Machine Learning experiments encapsulate a data flow in which tables of data are passed between modules as input parameters. Each module can then operate on the data, and pass the modified data as an output parameter to the next module in the data flow; as shown here:

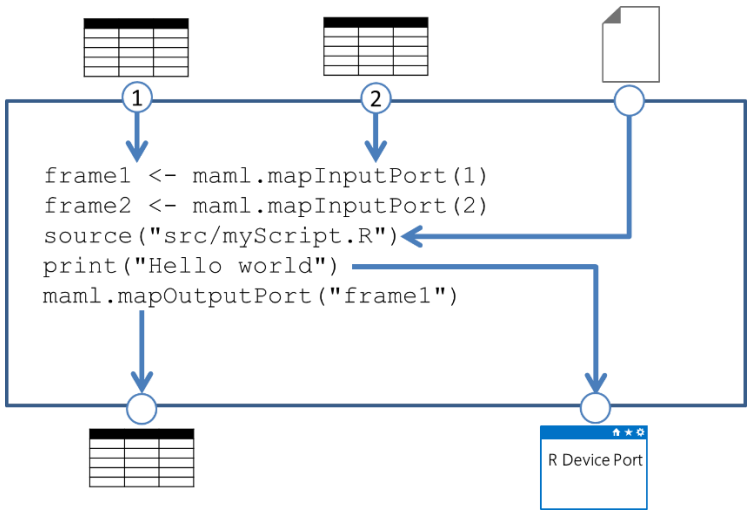


Data Frames in R

R provides native support for data frames, and includes a library named **dplyr** that provides an extensive set of operations for working with them. You can download the documentation for the dplyr library at https://cran.r-project.org/web/packages/dplyr/dplyr.pdf.

In an Azure Machine Learning experiment, you can use an **Execute R Script** module to run custom R code on one or two data frames, which are passed to the module as input parameters from the

experiment data flow. In your custom code, you can manipulate the data frames using standard R operations or **dplyr** functions, and you can pass a data frame as an output to the next module in the experiment data flow. You can also import custom R libraries that you have uploaded to Azure ML as R scripts in a zip file, and you can write values to the device output just as you would print to the console in an R IDE such as RStudio.



Data Frames in Python

Python supports data frames through the **pandas** library. You can view the documentation for the pandas library at http://pandas.pydata.org/pandas-docs/version/0.18.1/index.html.

In an Azure Machine Learning experiment, you can use an **Execute Python Script** module to run custom Python code on one or two data frames, which are passed to the module as input parameters from the experiment data flow. In your custom code, you can manipulate the data frames using **pandas** functions, and you can pass a data frame as an output to the next module in the experiment data flow. You can also import custom Python modules that you have uploaded to Azure ML as Python scripts in a zip file, and you can write values to the device output just as you would print to the console in a Python IDE such as Spyder.