

2.4: Data Frames

The fundamental object used for storing tables of data in R is called a data frame. We can think of a data frame as a way of organizing data into a large table with a row for each system measured and a column for each parameter. An interesting and useful feature of R is that all the columns in a data frame do not need to be the same data type. Some columns may consist of numerical data, for instance, while other columns contain textual data. This feature is quite useful when manipulating large, heterogeneous data files.

To access the CPU DB data, we first must read it into the R environment. R has built-in functions for reading data directly from files in the csv (comma separated values) format and for organizing the data into data frames. The specifics of this reading process can get a little messy, depending on how the data is organized in the file. We will defer the specifics of reading the CPU DB file into R until Chapter 6. For now, we will use a function called `extract_data()`, which was specifically written for reading the CPU DB file.

To use this function, copy both the `all-data.csv` and `read-data.R` files into a directory on your computer (you can download both of these files from this book's web site shown on p. ii). Then start the R environment and set the local directory in R to be this directory using the File -> Change dir pull-down menu. Then use the File -> Source R code pull-down menu to read the `read-data.R` file into R. When the R code in this file completes, you should have six new data frames in your R environment workspace: `int92.dat`, `fp92.dat`, `int95.dat`, `fp95.dat`, `int00.dat`, `fp00.dat`, `int06.dat`, and `fp06.dat`.

The data frame `int92.dat` contains the data from the CPU DB database for all of the processors for which performance results were available for the SPEC Integer 1992 (Int1992) benchmark program. Similarly, `fp92.dat` contains the data for the processors that executed the Floating-Point 1992 (Fp1992) benchmarks, and so on. I use the `.dat` suffix to show that the corresponding variable name is a data frame.

Simply typing the name of the data frame will cause R to print the entire table. For example, here are the first few lines printed after I type `int92.dat`, truncated to fit within the page: `nperf perf clock threads cores ... 1 9.662070 68.60000 100 1 1 ... 2 7.996196 63.10000 125 1 1 ... 3 16.363872 90.72647 166 1 1 ... 4 13.720745 82.00000 175 1 1 ...` The first row is the header, which shows the name of each column. Each subsequent row contains the data corresponding to an individual processor. The first column is the index number assigned to the processor whose data is in that row. The next columns are the specific values recorded for that parameter for each processor. The function `head(int92.dat)` prints out just the header and the first few rows of the corresponding data frame. It gives you a quick glance at the data frame when you interact with your data.

Table 2.1 shows the complete list of column names available in these data frames. Note that the column names are listed vertically in this table, simply to make them fit on the page.

Table 2.1: The names and definitions of the columns in the data frames containing the data from CPU DB.

Column number	Column name	Definition
1	(blank)	Processor index number
2	nperf	Normalized performance
3	perf	SPEC performance
4	clock	Clock frequency (MHz)
5	threads	Number of hardware threads available
6	cores	Number of hardware cores available
7	TDP	Thermal design power
8	transistors	Number of transistors on the chip (M)
9	dieSize	The size of the chip
10	voltage	Nominal operating voltage

11	featureSize	Fabrication feature size
12	channel	Fabrication channel size
13	FO4delay	Fan-out-four delay
14	L1icache	Level 1 instruction cache size
15	L1dcache	Level 1 data cache size
16	L2cache	Level 2 cache size
17	L3cache	Level 3 cache size

This page titled [2.4: Data Frames](#) is shared under a [CC BY-NC 4.0](#) license and was authored, remixed, and/or curated by [David Lilja](#) ([University of Minnesota Libraries Publishing](#)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.