Using SQL for data consolidation in R

Jennifer Vinas-Forcade[1], Julien Nacci, Cindy Mels[2], Martin Valcke[1], Ilse Derluyn[1]
[1] Ghent University, [2] Universidad Católica del Uruguay

Keywords: SQL, sqldf, R, database consolidation, data cleaning.

Working with multiple data sources implies data cleaning and consolidation prior to analysis. R has become popular among social scientists (Kelley, 2007; Clark, 2014), who are advised to screen data in a "favorite spreadsheet program" (Muenchen, 2011:21), before importing it to R. This way, users avoid typing in the R console and are supported by a graphical user interface. Even for experienced R users, querying/retrieving data from multiple large sources takes a lot of computing power, which is better handled by SQL language (Table 2; KeyCentrix, 2015).

Examples of the main commands of the R 'sqldf' package in Table 1. Differences between SQL and R languages in Table 2.

Table 1. SQL functions used in 'sqldf' for data cleaning and database consolidation

Task	Function(s)
Data cleaning: identify unique values	_Select_ **_distinct_** _... from ..._
Data cleaning: delete missing values	_Select… from ... where ..._ **_is not null_**
Merging data (union / add rows)	_Select ..._ **_union_** _select ..._ **_union_** _select ..._
Merge data frames with different # of columns	_Select df1.v1, df1.v2, df1.v3 from df1_ **_union_** _df2.v1,_ **_df2.null_**_, df2.v3 from df2_
Consolidate n data frames using unique id, discard all non-matches	_Select df1.v1, df2.v1 from_ **_df1, df2_** _where_ **_df1.id = df2.id_**
Consolidate n data frames keeping all baseline records	_Select df1.*, df2.* from df1_ **_left join_** _df2 on df1.id = df2.id_
Basic data aggregation operations	_Select ..._ **_count_** _(...),_ **_avg_** _(...)_ **_group by_** _..._
Data integrity (check-ups)	_Select ... where v1_ **_[not] in_** _(select ...)_
Reorder columns of a data frame	_Select_ **_v3, v4, v2, v1_** _from df_

Table 2. Differences between SQL and R languages.

	SQL	R
Function	Data optimizing, updating, querying	Statistical data analysis
Math&stats	Only basic operations	Specific functions for complex operations.
Syntax	More anthropomorphic	Less intelligible
Memory	Retrieves the specific data needed for each query, when prompted.	Loads all data on RAM memory.

Although SQL and R have similar toolsets, the nature of SQL and 'sqldf', make it more agile for data structuring and querying prior to data analysis with R.