Working with databases in R

Christopher Maronga - R Programmer

June 12, 2021

About me

- Training:
  - Bsc. Mathematics - JKUAT
  - MSc. Statistical Science - Strathmore University.
- Profession (1/2):
  - Over 7 years experience in health research data management and analysis.
  - Approx. 4 years in HIV clinical trials and reproductive health studies (KEMRI-CCR)
  - Designed over a dozen databases (REDCap & MySQL).
- Profession (2/2)
  - Data Manager (KEMRI-Wellcome Trust)
  - Malnutrition and mortality in children CHAIN
  - Worflow - REDCap, MySQL and Shiny Dashboards
- Interests
  - Joint modelling of longitudinal and survival data
  - Predictive modeling and data mining

The collaboration

Objectives

» Over 90% of this session will be a hands-on i.e. walking through code examples and implementation. The completed scripts will be shared after the training.

- What you will learn:-
  - Efficiently connecting R to RDBMS
  - Querying data stored in the RDBMS via R
  - Connecting and exporting data from REDCap
  - API security
- What won't be covered:-
  - Designing RDBMS
  - Database designing and survey authoring in REDCap
  - High level SQL scripting (we will substitute this with R code)

Introduction to RDBMS

- Storage and management of data is widely done using RDBMS
- Data is organized in tables(relations) with rows(tuples) and columns(variables)
- Tables are inter-related via specific attributes (unique keys)
- Majority of RDBMS use `SQL` to query and manage data
- Most popular RDBMS:-
  - MySQL, SQLite, PostgreSQL, Oracle, MariaDB, Ms Access etc.

Research Electronic Data Capture (REDCap)

- REDCap is a secure web application for bulding and managing online surveys and databases .
- Created at Vanderbilt University and can be used to collect virtually any kind of data:-
  - Supports offline data collection (REDCap mobile app)
  - Online surveys
- Used by over 5146 institutions, visit here for more information.

Connecting and querying a RDBMS

- Several ways to connect a RDBMS and R.
  - Use applicable `odbc` driver e.g. `MySQL ODBC 8.0 Unicode Driver`
  - `DBI` enabling package e.g. `RMySQL`, `RSQLite`, `ROracle` etc.
  - Use `SQL` code chunk in `Rmarkdown` file
- Ways to query RDBMS using R
  - Use DBI functions
  - Apply `dplyr` syntax (doesn't need knowledge of SQL)
  - Run `SQL` code chunks in `Rmarkdown` files

APIs Security

- This entails protecting the integrity of the API
- API credentials in the wrong hands is a data breach and can cost your organization a fortune
- Care should be taken to manage and secure API tokens and should not be put directly in your code
- In R, one such way is to keep API keys a secret is via `.Renviron` start-up file
- Warning: Do NOT share codes containing API keys. For GitHub, be sure to include `.Renviron` in `.gitignore` file.

Reference material

- Blog from RStudio, Databases in R
- Vebash Naidoo's blog posts; using the tidyverse with databases
- Modern Data Science with R: database querying
- R startup files
- Introduction to SQL
- SQL tutorial
- Setting up a database server