California Teachers Study ETL Methodology

Emma Spielfogel

2024-06-27

Table of contents

Preface **3**

1 Introduction **4**

2 RStudio Setup **5**

2.1 Connecting to data in s3 5

2.2 Using Git and GitHub 7

References **8**

Preface

This is a Quarto book for documenting the CTS' ETL methodology. This book is currently under development.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 RStudio Setup

This section will describe how to set up RStudio for development.

2.1 Connecting to data in s3

To connect from a local RStudio installation to s3 buckets, you need to:

1. Install AWS cli. Instructions can be found here.
2. Configure AWS sts single sign-on (SSO). Instructions can be found here.
3. Open a terminal window and type "aws sso login"--a browser window should open to allow you to log in to AWS.

> **You will need to follow step 3 and log in each time the session has expired.**

The paws package can be used for bringing data from s3 into your R environment. Another option is the aws.s3 package, which is not explored here.

The below code is an example of bringing data from an s3 bucket into RStudio for development.

```
# install paws if not already installed
install.packages(paws)

# load packages
library(paws)
library(tidyverse)
library(magrittr)

# create s3 frame to use paws package
s3 <- paws::s3()

# list buckets
s3$list_buckets()
```

```r
# list objects in a bucket
viewobs <- s3$list_objects(Bucket = "your-bucket-name")

# set variables for file you'd like to read in
bucket_name = "your-bucket-name"
file_name = "your-file-name.csv"

# download the file and store the output in a variable
s3_download <- s3$get_object(
  Bucket = bucket_name,
  Key = file_name
)

# Two options for consuming the data in RStudio

## Option 1 - save to local directory in RStudio

# read the CSV in from disk and write it to the current directory
file_name_save <- "s3_download.csv"
writeBin(s3_download$Body, con = file_name_save)

# read in as csv
file_in <- read.csv(file_name_save)

# code to remove file if undesired to keep locally
file.remove(file_name_save)

## Option 2 - read in the CSV directly from S3 - this requires magrittr package
require(magrittr)

# read in without writing to disk
file_in <- s3_download$Body %>%
  rawToChar %>%
  read.csv(text = .)

# another example - writing to data frame, using read_csv and reading all cols as char
file_in <- s3_download$Body %>%
  rawToChar %>%
  read_csv(col_types = cols(.default = "c"))
```

2.2 Using Git and GitHub

Follow the instructions here to set up Git and GitHub. As described in these instructions, using R Studio Projects makes working with Git and GitHub easier.

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.