# Using DuckDB with WRDS data

Ian D. Gow

22 December 2023

Demonstrate the power of DuckDB and dbplyr with WRDS data.

In this short note, I show how one can use DuckDB with WRDS data stored in the PostgreSQL database provided by WRDS. I then use some simple benchmarks to show how DuckDB offers a powerful, fast analytical engine for researchers in accounting and finance.

To make the analysis concrete, I focus on data used in the excellent recent book "Tidy Finance with R". Essentially, I combine data from CRSP's daily stock return file (`crsp.dsf`) with data on factor returns from Ken French's website and then run an aggregate query.

## 1 Summary of findings

While using DuckDB simplifies the process of collecting data from WRDS (and results in a shorter download time), the real differences come after the data sets are on your computer. Using DuckDB to load the data and run an aggregate query reduces the time taken from over two minutes using `dplyr` to well under one second. DuckDB from disk is faster than `dplyr` from RAM. Additionally DuckDB is faster than SQLite. In fact, for many queries DuckDB would be faster collecting data from WRDS than `dplyr` is with data in a local SQLite database. While performance isn't everything, gains like these likely deliver real quality-of-life benefits to data analysts.

I also show that almost all the performance benefits of DuckDB are realized even if the data are stored in parquet files. This is useful information because, while the format of DuckDB database files remains in flux, parquet files are regarded by many as the "modern CSV" and can be read by many software systems, including R and Python. I describe how one could maintain a local library of parquet files including copies of WRDS tables here.

This note illustrates the power of the core Tidy Finance approach. With a few tweaks, one can springboard from the SQLite-and-`dplyr` approach of the book to the very cutting-edge of data science tools and approaches.

## 2 Databases and tidy data

A popular way to manage and store data is with SQL databases. Tidy Finance with R uses SQLite, which "implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine." In this note, I use DuckDB, which has been described as offering "SQLite for Analytics". DuckDB is like SQLite in not requiring a server process, but like server-based databases such as PostgreSQL in terms of support for advanced SQL features and data types.

While storing data in a DuckDB database offers some benefits of SQLite (e.g., data compression), the real benefits of using DuckDB come from using the database engine for data analytics. For the most part, Tidy Finance with R uses SQLite for storage and uses `dplyr` and in-memory data frames for analysis. For example, in the chapter on beta estimation, the data are read into memory immediately using `collect()` before any analysis is conducted. However, the `dbplyr` package allows many analytical tasks to be performed in the database. In this note, I demonstrate how using DuckDB and `dbplyr` can lead to significant performance gains.

## 3 Getting data

There are two data sets that we need to collect. The first is the factor returns, which we collect from Ken French's website using the `frenchdata` package. The second is from CRSP's daily stock file, which we get from WRDS.

We start by loading three packages. Note that we load `DBI` rather than the underlying database driver package.[1] In addition to these three packages, you should have the `duckdb` and `RSQLite` packages installed. Use `install.packages()` in R to install any missing packages.

```
library(tidyverse)
library(DBI)
library(frenchdata)
library(arrow)
```

Next we set up a DuckDB database file in the `data` directory, creating this directory if need be. We set `read_only = FALSE` because we will want to write to this database connection.

```
if (!dir.exists("data")) dir.create("data")

tidy_finance <- dbConnect(
```

[1] This is how it's done in "R for Data Science". I have read comments by Hadley Wickham that this is the right way to do it, but I can't find those comments.

```
  duckdb::duckdb(),
  "data/tidy_finance.duckdb",
  read_only = FALSE)
```

### 3.1 Fama-French factor returns

We use the same `start_date` and `end_date` values used in "Tidy Finance with R" and the code below also is adapted from that book. However, we use the `copy_to()` function from `dplyr` to save the table to our database.

```
start_date <- ymd("1960-01-01")
end_date <- ymd("2021-12-31")

factors_ff_daily_raw <-
  download_french_data("Fama/French 3 Factors [Daily]")
```

```
New names:
* `` -> `...1`
```

```
factors_ff_daily <-
  factors_ff_daily_raw$subsets$data[[1]] |>
  transmute(
    date = ymd(date),
    rf = as.numeric(RF) / 100,
    mkt_excess = as.numeric(`Mkt-RF`) / 100,
    smb = as.numeric(SMB) / 100,
    hml = as.numeric(HML) / 100
  ) |>
  filter(date >= start_date & date <= end_date) |>
  copy_to(tidy_finance,
          df = _,
          name = "factors_ff_daily",
          temporary = FALSE,
          overwrite = TRUE)
```

### 3.2 Getting daily returns from WRDS

Next, I specify the connection details as follows. I recommend using environment variables (e.g., set using `Sys.setenv()`), as this facilitates sharing code with others. You should not

include this chunk of code in your code, rather run it before executing your other code. In addition to setting these environment variables, you may want to set `PGPASSWORD` too. (Hopefully it is obvious that your should use *your* WRDS ID and password, not mine.)

```
Sys.setenv(PGHOST = "wrds-pgdata.wharton.upenn.edu",
           PGPORT = 9737L,
           PGDATABASE = "wrds",
           PGUSER = "iangow")
```

Third, we connect to the CRSP daily stock file in the WRDS PostgreSQL database.

```
pg <- dbConnect(RPostgres::Postgres())
dsf_db <- tbl(pg, Id(schema = "crsp", table = "dsf"))
```

As we can see, we have access to data in `crsp.dsf`.

```
dsf_db
```

```
# Source:   table<dsf> [?? x 20]
# Database: postgres  [iangow@wrds-pgdata.wharton.upenn.edu:9737/wrds]
   cusip    permno permco issuno hexcd hsiccd date       bidlo askhi   prc   vol
   <chr>     <dbl>  <dbl>  <dbl> <dbl>  <dbl> <date>     <dbl> <dbl> <dbl> <dbl>
 1 68391610  10000   7952  10396     3   3990 1986-01-07  2.38  2.75 -2.56  1000
 2 68391610  10000   7952  10396     3   3990 1986-01-08  2.38  2.62 -2.5  12800
 3 68391610  10000   7952  10396     3   3990 1986-01-09  2.38  2.62 -2.5   1400
 4 68391610  10000   7952  10396     3   3990 1986-01-10  2.38  2.62 -2.5   8500
 5 68391610  10000   7952  10396     3   3990 1986-01-13  2.5   2.75 -2.62  5450
 6 68391610  10000   7952  10396     3   3990 1986-01-14  2.62  2.88 -2.75  2075
 7 68391610  10000   7952  10396     3   3990 1986-01-15  2.75  3    -2.88 22490
 8 68391610  10000   7952  10396     3   3990 1986-01-16  2.88  3.12 -3    10900
 9 68391610  10000   7952  10396     3   3990 1986-01-17  2.88  3.12 -3     8470
10 68391610  10000   7952  10396     3   3990 1986-01-20  2.88  3.12 -3     1000
# i more rows
# i 9 more variables: ret <dbl>, bid <dbl>, ask <dbl>, shrout <dbl>,
#   cfacpr <dbl>, cfacshr <dbl>, openprc <dbl>, numtrd <dbl>, retx <dbl>
```

Before proceeding with our first benchmark, we will make a version of `system.time()` that works with assignment.[2]

[2] If we put `system.time()` at the end of this pipe, then `crsp_daily` would hold the value returned by that function rather than the result of the pipeline preceding it. At first, the `system_time()` function may seem like magic, but Hadley Wickham explained to me that this works because of **lazy evaluation**, which is discussed in "Advanced R" here. Essentially, `x` is evaluated just once—inside `system.time()`—and its value is returned in the next line.

```
system_time <- function(x) {
  print(system.time(x))
  x
}
```

The following code is adapted from the Tidy Finance code here. But the original code is much more complicated and takes slightly longer to run.[3]

```
rs <- dbExecute(tidy_finance, "DROP TABLE IF EXISTS crsp_daily")

crsp_daily <-
  dsf_db |>
  filter(between(date, start_date, end_date),
         !is.na(ret)) |>
  select(permno, date, ret) |>
  mutate(month = as.Date(floor_date(date, "month"))) |>
  copy_to(tidy_finance, df = _, name = "dsf_temp") |>
  left_join(factors_ff_daily |>
              select(date, rf), by = "date") |>
  mutate(
    ret_excess = ret - rf,
    ret_excess = pmax(ret_excess, -1, na.rm = TRUE)
  ) |>
  select(permno, date, month, ret_excess) |>
  compute(name = "crsp_daily", temporary = FALSE, overwrite = TRUE) |>
  system_time()
```

```
   user  system elapsed
 92.876   9.689 214.094
```

### 3.3 Saving data to SQLite

If you have been working through "Tidy Finance", you may already have an SQLite database containing `crsp_daily`. If not, we can easily create one now and copy the table from our DuckDB database to SQLite.

[3] Performance will vary according to the speed of your connection to WRDS. Note that this query does temporarily use a significant amount of RAM on my machine, it is not clear that DuckDB will use as much RAM if this is more constrained. If necessary, you can run (say) `dbExecute(tidy_finance, "SET memory_limit='1GB'")` to constrain DuckDB's memory usage; doing so has little impact on performance for this query.

```
tidy_finance_sqlite <- dbConnect(
  RSQLite::SQLite(),
  "data/tidy_finance.sqlite",
  extended_types = TRUE)
```

```
copy_to(tidy_finance_sqlite,
        crsp_daily,
        name = "crsp_daily",
        overwrite = TRUE,
        temporary = FALSE)

dbExecute(tidy_finance_sqlite, "VACUUM")
```

We can also save the data to a parquet file.

```
dbExecute(tidy_finance,
          "COPY crsp_daily TO 'data/crsp_daily.parquet'
          (FORMAT 'PARQUET')")
```

Having created our two databases, we disconnect from them. This mimics the most common "write-once, read-many" pattern for using databases.

```
dbDisconnect(tidy_finance_sqlite)
dbDisconnect(tidy_finance, shutdown = TRUE)
```

## 4 Benchmarking a simple aggregation query

The following is a simple comparison of several different ways of doing some basic data analysis with R. After running the code above, we have the table `crsp_daily` as described in Tidy Finance in two separate databases—a SQLite database and a DuckDB database—and in a parquet file.

The following examines the same query processed in three different ways.

1. Using `dplyr` on an in-memory dataframe
2. Using `dbplyr` with an SQLite database
3. Using `dbplyr` with a DuckDB database
4. Using `dbplyr` with DuckDB and a parquet file.
5. Using `dbplyr` with the `arrow` library and a parquet file.

### 4.1 dplyr

We first need to load the data into memory.

```
tidy_finance <- dbConnect(
  RSQLite::SQLite(),
  "data/tidy_finance.sqlite",
  extended_types = TRUE)

crsp_daily <- tbl(tidy_finance, "crsp_daily")
```

What takes most time is simply loading nearly 2GB of data into memory.

```
crsp_daily_local <-
  crsp_daily |>
  collect() |>
  system_time()
```

```
   user  system elapsed
115.481   1.833 118.783
```

Once the data are in memory, it is *relatively* quick to run a summary query.

```
crsp_daily_local |>
  group_by(month) |>
  summarize(ret = mean(ret_excess, na.rm = TRUE)) |>
  collect() |>
  system_time()
```

```
   user  system elapsed
  3.035   0.597   3.751

# A tibble: 744 x 2
   month            ret
   <date>         <dbl>
 1 1960-01-01 -0.00213
 2 1960-02-01  0.000325
 3 1960-03-01 -0.00115
 4 1960-04-01 -0.00106
 5 1960-05-01  0.00114
```

```
 6 1960-06-01  0.000935
 7 1960-07-01 -0.000955
 8 1960-08-01  0.00159
 9 1960-09-01 -0.00289
10 1960-10-01 -0.00116
# i 734 more rows
```

```
rm(crsp_daily_local)
```

### 4.2 dbplyr with SQLite

Things are faster with SQLite, though there's no obvious way to split the time between reading the data and performing the aggregation. Note that we have a `collect()` at the end. This will not take a noticeable amount of time, but seems to be a reasonable step if our plan is to analyse the aggregated data in R.

```
crsp_daily |>
  group_by(month) |>
  summarize(ret = mean(ret_excess, na.rm = TRUE)) |>
  collect() |>
  system_time()
```

```
   user  system elapsed
 20.329   3.440  31.739
```

```
# A tibble: 744 x 2
   month             ret
   <date>          <dbl>
 1 1960-01-01 -0.00213
 2 1960-02-01  0.000325
 3 1960-03-01 -0.00115
 4 1960-04-01 -0.00106
 5 1960-05-01  0.00114
 6 1960-06-01  0.000935
 7 1960-07-01 -0.000955
 8 1960-08-01  0.00159
 9 1960-09-01 -0.00289
10 1960-10-01 -0.00116
# i 734 more rows
```

```
dbDisconnect(tidy_finance)
```

### 4.3 dbplyr with DuckDB

Let's consider DuckDB. Note that we are only reading the data here, so we set `read_only = TRUE` in connecting to the database. Apart from the connection, there is no difference between the code here and the code above using SQLite.

```
tidy_finance <- dbConnect(
  duckdb::duckdb(),
  "data/tidy_finance.duckdb",
  read_only = TRUE)
```

```
crsp_daily <- tbl(tidy_finance, "crsp_daily")
```

```
crsp_daily |>
  group_by(month) |>
  summarize(ret = mean(ret_excess, na.rm = TRUE)) |>
  collect() |>
  system_time()
```

```
   user  system elapsed
  1.695   0.298   0.372
```

```
# A tibble: 744 x 2
   month             ret
   <date>          <dbl>
 1 1991-08-01  0.00171
 2 1992-06-01 -0.000759
 3 1992-11-01  0.00466
 4 1992-12-01  0.00298
 5 1993-04-01  0.000248
 6 1993-06-01  0.00127
 7 1993-07-01  0.00141
 8 1987-02-01  0.00370
 9 1987-07-01  0.00150
10 1988-11-01 -0.00161
# i 734 more rows
```

Having done our benchmarks, we can take a quick peek at the data.

```
crsp_daily |>
  group_by(month) |>
  summarize(ret = mean(ret_excess, na.rm = TRUE)) |>
  arrange(month) |>
  collect()
```

```
# A tibble: 744 x 2
   month             ret
   <date>          <dbl>
 1 1960-01-01 -0.00213
 2 1960-02-01  0.000325
 3 1960-03-01 -0.00115
 4 1960-04-01 -0.00106
 5 1960-05-01  0.00114
 6 1960-06-01  0.000935
 7 1960-07-01 -0.000955
 8 1960-08-01  0.00159
 9 1960-09-01 -0.00289
10 1960-10-01 -0.00116
# i 734 more rows
```

Finally, we disconnect from the database. This will happen automatically if we close R, etc., and is less important if we have `read_only = TRUE` (so there is no lock on the file), but we keep things tidy here.

```
dbDisconnect(tidy_finance, shutdown = TRUE)
```

### 4.4 dbplyr with DuckDB and a parquet file

Let's do the benchmark using the parquet data.

```
db <- dbConnect(duckdb::duckdb())
crsp_daily <- tbl(db, "read_parquet('data/crsp_daily.parquet')")
```

```
crsp_daily |>
  group_by(month) |>
  summarize(ret = mean(ret_excess, na.rm = TRUE)) |>
  collect() |>
```

```
  system_time()
```

```
   user  system elapsed
  1.984   0.371   0.386
```

```
# A tibble: 744 x 2
   month             ret
   <date>          <dbl>
 1 1990-11-01  0.00258
 2 1991-04-01  0.00176
 3 1992-04-01 -0.00114
 4 1993-12-01  0.00153
 5 1994-01-01  0.00285
 6 1994-02-01  0.0000257
 7 1989-04-01  0.00135
 8 1990-06-01  0.000423
 9 1994-05-01  0.000753
10 1995-06-01  0.00272
# i 734 more rows
```

```
dbDisconnect(db, shutdown = TRUE)
```

### 4.5 The `arrow` library with a pqrquet file

Let's do one more benchmark using the parquet data with the `arrow` library.

```
crsp_daily <- open_dataset('data/crsp_daily.parquet')
```

```
crsp_daily |>
  group_by(month) |>
  summarize(ret = mean(ret_excess, na.rm = TRUE)) |>
  collect() |>
  system_time()
```

```
   user  system elapsed
  1.403   0.205   0.253
```

```
# A tibble: 744 x 2
   month            ret
   <date>         <dbl>
 1 1995-12-01  0.000953
 2 1996-01-01  0.00200
 3 1996-02-01  0.00188
 4 1996-03-01  0.00148
 5 1996-04-01  0.00291
 6 1996-05-01  0.00303
 7 1996-06-01 -0.00137
 8 1996-07-01 -0.00332
 9 1996-08-01  0.00239
10 1996-09-01  0.00168
# i 734 more rows
```

### 4.6 Python and parquet files

It seems that Python (with `pandas`) is constrained by the need to load the data into memory.

```python
import pandas as pd
import time
start_time = time.monotonic()
df = pd.read_parquet('data/crsp_daily.parquet')
avg = df.groupby('month').agg(ret = ('ret_excess', 'mean'))
seconds = time.monotonic() - start_time
print("Time Taken: {:0.2f} seconds.".format(seconds))
print(avg)
```

```
Time Taken: 19.07 seconds.
                ret
month
1960-01-01 -0.002130
1960-02-01  0.000325
1960-03-01 -0.001154
1960-04-01 -0.001062
1960-05-01  0.001137
...              ...
2021-08-01  0.000985
2021-09-01 -0.001301
2021-10-01  0.001344
2021-11-01 -0.002161
```

```
2021-12-01  0.000075

[744 rows x 1 columns]
```

But using the PyArrow library, we see performance very similar to that using R above. Of course, DuckDB offers an API for Python too, so parquet files offer the data analyst many ways of working with them.

```python
import pyarrow.parquet as pq
import time
start_time = time.monotonic()
df = pq.read_table('data/crsp_daily.parquet')
avg = df.group_by('month').aggregate([('ret_excess', 'mean')])
avg = avg.to_pandas().sort_values(by=['month'])
seconds = time.monotonic() - start_time
print("Time Taken: {:0.2f} seconds.".format(seconds))
print(avg)
```

```
Time Taken: 0.41 seconds.
          month  ret_excess_mean
0    1960-01-01        -0.002130
1    1960-02-01         0.000325
2    1960-03-01        -0.001154
3    1960-04-01        -0.001062
4    1960-05-01         0.001137
..          ...              ...
720  2021-08-01         0.000985
721  2021-09-01        -0.001301
722  2021-10-01         0.001344
723  2021-11-01        -0.002161
724  2021-12-01         0.000075

[744 rows x 2 columns]
```