D3

Nicolas Palombo

Neste documento será apresentado como é possível carregar arquivos em formato distinto do csv, formato esse mais usado. O banco de dados utilizado foi o “Motor Trend Car Road Tests” em JSONe o banco de dados “House Prices” para Parquet. Ambos datasets foram tirados do site: https://www.tablab.app

Carregando as bibliotecas necessárias:

```r
library(arrow)
```

```
Warning: pacote 'arrow' foi compilado no R versão 4.5.1
```

```

Anexando pacote: 'arrow'
```

```
O seguinte objeto é mascarado por 'package:utils':

    timestamp
```

```r
library(dplyr)
```

```
Warning: pacote 'dplyr' foi compilado no R versão 4.5.1
```

```

Anexando pacote: 'dplyr'
```

```
Os seguintes objetos são mascarados por 'package:stats':

    filter, lag

Os seguintes objetos são mascarados por 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(jsonlite)
```

```
Warning: pacote 'jsonlite' foi compilado no R versão 4.5.1
```

```r
library(rjson)
```

```


Anexando pacote: 'rjson'

Os seguintes objetos são mascarados por 'package:jsonlite':

    fromJSON, toJSON
```

```r
library(bench)
```

```
Warning: pacote 'bench' foi compilado no R versão 4.5.1
```

```r
library(ndjson)
```

```
Warning: pacote 'ndjson' foi compilado no R versão 4.5.1


Anexando pacote: 'ndjson'

Os seguintes objetos são mascarados por 'package:jsonlite':

    flatten, stream_in, validate
```

Explicação dos pacotes utilizados:

Arrow: O pacote *arrow* fornece acesso a muitos dos recursos da biblioteca Apache Arrow em C++. Entre suas diversas funcionalidades, destacam-se a leitura e escrita de arquivos no formato Parquet.

Jsonlite: Um analisador e gerador de JSON rápido e otimizado para dados estatísticos, oferecendo ferramentas simples e flexíveis para trabalhar com JSON em R.

NDJSON: Permite a leitura de objetos NDJSON ou NDJSON comprimidos

Bench: Fornece ferramentas para analisar com precisão a execução de expressões em R.

Principais colunas do bench:

expression -> código que foi medido

min -> menor tempo de execução

median -> mediana dos tempos de execução

itr/sec -> número de iterações por segundo

mem_alloc -> memória alocada durante a execução

gc/sec -> quantas vezes o garbage collector foi chamado por segundo (quanto mais alto o valor maior o esforço de memória)

Abrindo o parquet

```
house_price_df <- read_parquet("~/dados/house-price.parquet")
head(house_price_df)
```

```
# A tibble: 6 x 13
     price  area bedrooms bathrooms stories mainroad guestroom basement
     <int> <int>    <int>     <int>   <int> <chr>    <chr>     <chr>   
1 13300000  7420        4         2       3 yes      no        no      
2 12250000  8960        4         4       4 yes      no        no      
3 12250000  9960        3         2       2 yes      no        yes     
4 12215000  7500        4         2       2 yes      no        yes     
5 11410000  7420        4         1       2 yes      yes       yes     
6 10850000  7500        3         3       1 yes      no        yes     
# i 5 more variables: hotwaterheating <chr>, airconditioning <chr>,
#   parking <int>, prefarea <chr>, furnishingstatus <chr>
```

```
bench::mark(
  arrow =read_parquet("~/dados/house-price.parquet")
)
```

```
# A tibble: 1 x 6
  expression      min   median `itr/sec` mem_alloc `gc/sec`
  <bch:expr> <bch:tm> <bch:tm>     <dbl> <bch:byt>    <dbl>
1 arrow        2.79ms   3.04ms      323.      21KB     22.3
```

Como foram usados 2 métodos para abrir o arquivo .json, primeiramente será testado qual o método mais rápido

```r
res <- bench::mark(
  jsonlite =
     invisible(capture.output(jsonlite::stream_in(file("~/dados/mtcars-parquet.json"))))
  ,
  ndjson   =
     invisible(capture.output(ndjson::stream_in("~/dados/mtcars-parquet.json")))
  ,
  check = FALSE
)
```

```
opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.
```

```
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
```

```
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
```

```
closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.
```

```
opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.
```

```
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
```

```
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
closing file input connection.
opening file input connection.
```

```
closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.
```

```
opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.
```

```
closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.
```

```
opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.
```

```
closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.
```

```
opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.
```

```
closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.
```

```
opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.

opening file input connection.

closing file input connection.
```

```
res
```

```
# A tibble: 2 x 6
  expression      min   median `itr/sec` mem_alloc `gc/sec`
  <bch:expr> <bch:tm> <bch:tm>     <dbl> <bch:byt>    <dbl>
1 jsonlite     3.37ms   3.69ms      266.  385.96KB     4.09
2 ndjson       3.03ms   3.26ms      299.    3.47MB     4.09
```

Visto que ndjson é mais rápido do que jsonlite faremos uma visualização do arquivo usando o ndjson

```
mtcars_df <- ndjson::stream_in("~/dados/mtcars-parquet.json")
head(mtcars_df)
```

```
      am  carb   cyl  disp  drat  gear    hp             model   mpg  qsec
   <num> <num> <num> <num> <num> <num> <num>            <char> <num> <num>
1:     1     4     6   160  3.90     4   110         Mazda RX4  21.0 16.46
2:     1     4     6   160  3.90     4   110     Mazda RX4 Wag  21.0 17.02
3:     1     1     4   108  3.85     4    93        Datsun 710  22.8 18.61
4:     0     1     6   258  3.08     3   110    Hornet 4 Drive  21.4 19.44
5:     0     2     8   360  3.15     3   175 Hornet Sportabout  18.7 17.02
6:     0     1     6   225  2.76     3   105           Valiant  18.1 20.22
      vs    wt
   <num> <num>
1:     0 2.620
2:     0 2.875
3:     1 2.320
4:     1 3.215
5:     0 3.440
6:     1 3.460
```