Filtrando dados

Miguel Carvalho Nascimento

Comparação de Desempenho

Carregando pacotes necessários:

```
# Carregando pacotes
library(dplyr)
library(data.table)
library(microbenchmark)
```

Criando objeto para testes de desempenho:

```
# Dados simulados
df <- data.frame(a = rnorm(1e6), b = rnorm(1e6))
dim(df)
```

```
[1] 1000000       2
```

Exemplo com "r-base"

```
# Método clássico com base R
system.time({
  res_base <- df[df$a > 0 & df$b > 0, ]
})
```

```
##   usuário   sistema decorrido
##     0.027     0.000     0.027
```

Exemplo com “dplyr”

```
        # Método otimizado com dplyr
        system.time({
          res_dplyr <- df %>%
            filter(a > 0, b > 0)
        })
```

```
  usuário   sistema decorrido
    0.016     0.007     0.022
```

Exemplo com “data.table”

```
        # Método otimizado com data.table
        dt <- as.data.table(df)
        system.time({
          res_dt <- dt[a > 0 & b > 0]
        })
```

```
  usuário   sistema decorrido
    0.018     0.001     0.018
```

Comparação

```
        microbenchmark(
          base = df[df$a > 0 & df$b > 0, ],
          dplyr = df %>% filter(a > 0, b > 0),
          data_table = dt[a > 0 & b > 0]
        )
```

```
Unit: milliseconds
       expr      min       lq     mean   median       uq      max neval cld
       base 21.02837 23.20559 24.92932 23.97307 25.05379 63.83783   100  a
      dplyr 13.07921 14.31550 17.16939 15.43126 17.06880 53.68936   100   b
 data_table 12.46348 12.96692 15.12743 13.99795 15.59528 51.85688   100   b
```

Colunas Explicadas

- **Unit:** microseconds: A unidade de tempo usada para medir as expressões é microsegundos (1 microsegundo = 10^-6 segundos).
- **expr:** A expressão sendo avaliada.
- **min:** O menor tempo de execução observado entre todas as execuções.
- **lq (lower quartile):** O valor do primeiro quartil, ou seja, 25% das execuções foram concluídas em menos tempo que este valor.
- **mean:** A média dos tempos de execução.
- **median:** O valor mediano dos tempos de execução, ou seja, 50% das execuções foram concluídas em menos tempo que este valor.
- **uq (upper quartile):** O valor do terceiro quartil, ou seja, 75% das execuções foram concluídas em menos tempo que este valor.
- **max:** O maior tempo de execução observado entre todas as execuções.
- **neval:** O número de execuções realizadas para cada expressão.