



Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

02 de agosto de 2022

Lista 2: Manipulação em Bancos de dados e em Spark com R

Computação em Estatística para dados e cálculos massivos

Tópicos especiais em Estatística 1

Prof. Guilherme Rodrigues

César Augusto Fernandes Galvão (aluno colaborador)

Gabriel Jose dos Reis Carvalho (aluno colaborador)

José Vítor Barreto Porfírio 190089971 (**autor das soluções**)

1. As questões deverão ser respondidas em um único relatório *PDF* ou *html*, produzido usando as funcionalidades do *Rmarkdown* ou outra ferramenta equivalente.
2. O aluno poderá consultar materiais relevantes disponíveis na internet, tais como livros, *blogs* e artigos.
3. O trabalho é individual. Suspeitas de plágio e compartilhamento de soluções serão tratadas com rigor.
4. Os códigos *R* utilizados devem ser disponibilizados na íntegra, seja no corpo do texto ou como anexo.
5. O aluno deverá enviar o trabalho até a data especificada na plataforma Microsoft Teams.
6. O trabalho será avaliado considerando o nível de qualidade do relatório, o que inclui a precisão das respostas, a pertinência das soluções encontradas, a formatação adotada, dentre outros aspectos correlatos.
7. Escreva seu código com esmero, evitando operações redundantes, visando eficiência computacional, otimizando o uso de memória, comentando os resultados e usando as melhores práticas em programação.

Por vezes, mesmo fazendo seleção de colunas e filtragem de linhas, o tamanho final da tabela extrapola o espaço disponível na memória RAM. Nesses casos, precisamos realizar as operações de manipulação fora do R, em um banco de dados ou em um sistema de armazenamento distribuído. Outras vezes, os dados já estão armazenados em algum servidor/cluster e queremos carregar para o R parte dele, possivelmente após algumas manipulações. Nessa lista repetiremos parte do que fizemos na Lista 1. Se desejar, use o gabarito da Lista 1 em substituição à sua própria solução dos respectivos itens.

Preparação: Leitura dos bancos

Relembrando a solução do item 2a) da Lista 1, o objetivo era carregar para o ambiente de R os dados de vacinação que haviam sido baixados da base de dados do governo. Dessa vez seguiu-se carregando os arquivos pelo `vroom` por ter sido aparentemente mais rápido e poder usar o vetor de arquivos como parâmetro da função.

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load("vroom", "glue", "geobr", "rjson", "dplyr")

arquivos <- list.files("./Lista_1\\dados", full.names = TRUE)
vax_all <- vroom(
  arquivos,
  col_select = c(
    "estabelecimento_uf",
    "vacina_descricao_dose",
    "estabelecimento_municipio_codigo"
  ),
  show_col_types = FALSE,
  col_types = list("estabelecimento_municipio_codigo" = col_character()),
  num_threads = 7
)
```

Utilizando as facilidades do `geobr` e carregando o arquivo de códigos também baixado durante a Lista 1, dessa vez sem depender do `geobr`, pois o arquivo de códigos já contém todas as informações necessárias. Vale notar que ao poupar todas as soluções desse procedimento de *join* dos dados `geobr` com os códigos do IBGE o tempo poupado é mínimo, uma vez que os dois bancos são “pequenos”.

```
codigos <- vroom(
  "./Lista_2_gabarito/Tabela_codigos.csv",
  col_types = list("Cód IBGE" = col_character())
) %>%
  select(c("Cód IBGE", "Nome da Região de Saúde")) %>%
  rename_all(~ c(
    "est_mun_codigo", "name_health_region"
  ))
```

Questão 1: Criando bancos de dados.

a) Crie um banco de dados SQLite e adicione a tabela gerada no item 2a) da Lista 1.

Solução: Dispondo dos pacotes `DBI` e `RSQLite`, para criar o banco de dados basta inicializar a conexão com a `db` (*database*) pela função `dbConnect`, que por sua vez já lida com criar o banco caso ele ainda não exista. A função `dbWriteTable` consegue enviar um `data.frame`, ou similar, do R para a `db` como no código a seguir

```
sql_conn <- dbConnect(RSQLite::SQLite(), "./Lista_2_gabarito/lista2.sqlite")
dbWriteTable(sql_conn, "vax", vax_all)
dbWriteTable(sql_conn, "codigos", codigos)
```

b) Refaça as operações descritas no item 2b) da Lista 1 executando códigos `sql` diretamente no banco de dados criado no item a). Ao final, importe a tabela resultante para R. Não é necessário descrever novamente o que são as regiões de saúde.

Atenção: Pesquise e elabore os comandos `sql` sem usar a ferramenta de tradução de `dplyr` para `sql`.

Solução: Sabendo da existencia dessa tabela de vacinações, é possível utilizar uma *query* SQL para filtrar apenas as observações para as quais foi aplicada a segunda dose da vacina contidas na tabela criada no item a) com um simples `WHERE`, isto é,

```
SELECT * FROM vax
WHERE vacina_descricao_dose='1ª Dose'
```

Note no código fonte ao final do arquivo que a *query* não foi prontamente executada na base de dados, apenas criou-se uma `string` contendo a *query* para concatenar com as demais e executar a *query* em uma “tacada” só. Para as próximas *queries* vale lembrar que a notação “{variável}” é a notação do pacote `glue` para avaliar variáveis e inserir seu valor na *string*, mas ao longo das *queries* apresentadas neste item é apenas para se referir a como a *query* anterior foi alocada.

A implementação do `LEFT JOIN` em `SQLite` poderia ser feita para dar sequência à solução, toda via, concatenar os codigos de municipio dessa forma deixou a desejar em performance. Haja vista esse problema, a abordagem utilizada foi selecionar as observações do nome da região de saúde no banco de códigos que corresponde aos códigos dos estabelecimento no banco de vacinações e com isso criar uma coluna de saída

```
SELECT (
  SELECT name_health_region
  FROM codigos
  WHERE (codigos.estabelecimento_municipio_codigo =
        prim_dose.estabelecimento_municipio_codigo)
) AS regioao_saude
FROM ({last_query}) AS prim_dose
```

Segue-se do resultado da *query* anterior, e em seguida utilizando a função `COUNT(*)` do SQL em conjunto com um `GROUP BY` para contar o número de vacinados por região de saúde.

```
SELECT regioao_saude, COUNT(*) AS N
FROM ({last_query})
GROUP BY regioao_saude
```

Dispondo da frequência de cada uma das regiões de saúde no banco de vacinados, vamos utilizar a tabela gerada pela *query* anterior como “`qnt_vax`”, tomando cuidado com o posicionamento dos *statements* `WITH` e `AS`. Então, o procedimento para criar as faixas de vacinação em `SQLite` é através da operação de `CASE WHEN`, que funciona similarmente a um *if statement* do R

```
WITH qnt_vax AS ({last_query})
SELECT regioao_saude, N,
  CASE WHEN N > (SELECT MEDIAN(N) FROM qnt_vax)
  THEN 'Alto'
  ELSE 'Baixo'
  END AS Faixa
FROM qnt_vax
```

Note que a tabela filtrada pela faixa de vacinação e organizada pela coluna `N`, apresenta nas 5 primeiras observações as 5 menores observações de cada grupo. Agora basta avaliar essas *queries* e tomar a sua `UNION` como resultado, uma vez que possuem as mesmas colunas (equivalente a um `rbind` do R).

```

WITH tabFaixa AS ({last_query})
SELECT *
FROM (
  SELECT *
  FROM tabFaixa
  WHERE Faixa='Baixo'
  ORDER BY N ASC LIMIT 5)
UNION
SELECT *
FROM (
  SELECT *
  FROM tabFaixa
  WHERE Faixa='Alto'
  ORDER BY N ASC LIMIT 5)
ORDER BY N

```

c) Refaça os itens a) e b), agora com um banco de dados MongoDB.

Solução: De forma análoga aos itens a) e b), o pacote `mongolite` permite a conexão entre o ambiente do R e bases de dados MongoDB. A função `mongo` abre a conexão com a `db` e cria a coleção (tabela) caso necessário, já o método `$insert` envia dados em formato JSON através da conexão, tendo algumas facilidades para enviar diretamente um `data.frame` do R sem a necessidade de primeiro converter o objeto para JSON.

```

mongo_local <- function(collection) {
  conn <- mongo(
    collection = collection,
    db = "Lista2_db",
    url = "mongodb://localhost"
  )
  return(conn)
}

conn_vax <- mongo_local("vax")
conn_vax$insert(vax_all)

conn_codigos <- mongo_local("codigos")
conn_codigos$insert(codigos)

```

Para realizar uma sequência de operações é conveniente utilizar o método `$aggregate`, método que utiliza uma pipeline escrita em uma extensão de sintaxe para JSON para executar vários comandos em sequência, de forma que o resultado do anterior é levado para a execução do próximo método. O método `$lookup` faz o equivalente a um LEFT JOIN do SQL, o `$project` é semelhante ao `$find` e nesse caso escolhe quais colunas devem ser retornadas (assim como o `select` do `dplyr`), o `$match` atua de forma semelhante ao `filter` do `dplyr`, `$group` é um pouco mais *tricky* e atua como um `group by` juntamente a outros operadores (nesse caso o operador `$sum`, que está contando observações).

```

qnt_vax <- conn_vax$aggregate('[
  {
    "$lookup":
    {
      "from": "codigos",
      "localField": "estabelecimento_municipio_codigo",
      "foreignField": "est_mun_codigo",
      "as": "regDocs"
    }
  },

```

```

    {
      "$project":{
        "vacina_descricao_dose": 1,
        "estabelecimento_uf": 1,
        "name_health_region": "$regDocs.name_health_region"
      }
    },
    { "$match": { "vacina_descricao_dose": "1ª Dose" } },
    { "$group": {
      "_id": "$name_health_region",
      "N": {"$sum": 1}
    }
  }
]') %>%
  rename("Nome" = `_id`)

```

Diferente do SQL, optou-se por não seguir de forma direta em uma única *query*, mas sim por intrromper o procedimento por aqui para trazer o banco para o ambiente do R e computar a mediana do número de observações, uma vez que não há implementação “base” e nem trivial do cálculo da mediana em MongoDB.

```

mediana <- median(qnt_vax$N)

conn_qnt_vax <- mongo_local("qnt_vax")
qnt_vax %>%
  rename_all(~ c("reg_saude", "N")) %>%
  conn_qnt_vax$insert()

```

Com o banco contendo a quantidade de vacinados no mongo e dispondo da mediana, seguiu-se tomando cuidado com a notação do pacote **glue** para colocar “{}” reais e {} que chamam variável, isto é,

```

conn_qnt_vax$update(
  query = "{}",
  update = glue(' [{
    "$set": {
      "Faixa": {
        "$switch": {
          "branches": [
            { "case":
              {"$gte": ["$N", {mediana}]}},
              "then": "Alto" }},
            { "case":
              {"$lt": ["$N", {mediana}]}},
              "then": "Baixo" }
          ]
        }
      }
    }
  ]' ),
  multiple = TRUE,
  upsert = TRUE
)

```

Por fim, tomou-se o método **\$find** para encontrar as 5 primeiras observações de cada grupo, ordenada-las por N, e com ajuda da função **bind_rows** obteve-se um resultado análogo ao encontrado na Lista 1.

```
n <- 5
bot5 <- bind_rows(
  conn_qnt_vax$find(
    limit = n, sort = '{"N": 1}'
  ),
  conn_qnt_vax$find(
    skip = conn_qnt_vax$count() - n,
    limit = n, sort = '{"N": 1}'
  )
)
```

d) Refaça os itens c), agora usando o Apache Spark.

Solução: O primeiro passo é abrir a conexão com o Spark, que pode ser feita pelo pacote `sparklyr`, valendo notar que o banco de dados de vacinações a ser manipulado é especialmente grande e portanto utilizou-se das configurações da conexão Spark para aumentar o limite de memória para 16 *Gigabytes* (sem isso o Spark acusaria um erro no carregamento alegando falta de memória em *heap*), tudo como no código a seguir

```
conf <- spark_config()
conf$`sparklyr.shell.driver-memory` <- "16G"
conf$spark.memory.fraction <- 0.9
sc <- spark_connect(
  master = "local",
  version = "2.4.3",
  config = conf
)
```

A partir desse ponto há diversas formas de subir os dados para o Spark, pode-se usar a função `copy_to` para enviar um `dataframe` do R

```
tbl_cod <- copy_to(sc, codigos, name = "codigos", overwrite = TRUE)
```

porém esse método tem limitações para enviar dados muito grandes e após tentar enviar o banco carregado e enfrentar mensagens de erro, foi necessária uma mudança de método. É possível enviar os dados pelo `spark_read_csv` em uma chamada do `purrr::map`, ou mais simples ainda ao informar a pasta e os arquivos a serem lidos pelo *wildcard* `"*"`, uma vez que o `sparklyr` apresentou erro ao tentar passar um vetor com o caminho para os arquivos de uma só vez. Uma consequência de escolher o método que utilizava `map` é ter que mandar uma *query* para o Spark fazendo a união dos bancos lidos https://github.com/Voz-bonita/Big-Data-Compute/blob/master/Lista_2/spark.r.

```
tbl_vax <- spark_read_csv(sc = sc, path = "Lista_1/dados/*", delimiter = ";")
```

Lembrando que o Spark aceita *queries* escritas em SQL, é intuitivo pensar em utilizar as *queries* do item b) com algumas adaptações, isto é, vamos passar o banco de vacinações lido para a *query* SQL que no item b) filtrava a tabela “vax”, mas qual o nome da tabela que criamos em Spark anteriormente? Essa é uma pergunta da qual imagina-se que a resposta não seja previsível, não obstante, a API do Spark SQL dispõe do padrão `__THIS__` que aponta para a tabela atual

```
SELECT estabelecimento_uf,
       vacina_descricao_dose,
       estabelecimento_municipio_codigo
FROM __THIS__
WHERE vacina_descricao_dose='1ª Dose'
```

No item b) havia um problema de desempenho com o `LEFT JOIN`, mas agora a situação é outra. O procedimento de `JOIN` não será feito pelo SQL mesmo passando a *query* em SQL para o Spark,

ela será traduzida e então executada, portanto o melhor jeito de informar esse procedimento é com o próprio LEFT JOIN

```
SELECT name_health_region as regioao_saude
FROM ({last_query}) as prim_dose
LEFT JOIN codigos
ON prim_dose.estabelecimento_municipio_codigo
= codigos.est_mun_codigo
```

Para calcular a mediana teve-se outra mudança: a função MEDIAN não é corretamente traduzida para o Spark, o que geraria um erro. O problema pode ser resolvido utilizando uma função nativa do Spark, nesse caso a função percentile_approx que calcula os percentis solicitados

```
SELECT regioao_saude, COUNT(*) AS N
FROM ({last_query})
GROUP BY regioao_saude

WITH qntVax AS ({last_query})
SELECT regioao_saude, N,
CASE WHEN N > (
SELECT percentile_approx(N, 0.5)
FROM qntVax)
THEN 'Alto'
ELSE 'Baixo'
END AS Faixa
FROM qntVax
```

Também há funções mais adequadas do Spark para permitir encontrar os casos em que houveram menos vacinados por faixa de vacinação, nesse caso a função dense_rank em conjunto com a window function do SQL, como sugerido em <https://stackoverflow.com/questions/36660625/spark-sql-top-n-per-group>.

```
WITH tabFaixa AS ({last_query})
SELECT regioao_saude, N, Faixa
FROM (
SELECT *, dense_rank()
OVER (PARTITION BY Faixa ORDER BY N DESC) as posicao
FROM tabFaixa)
WHERE posicao <= 5
```

Agora é questão de enviar essas queries para o Spark através da função ft_sql_transformer e executar um collect para trazer os resultados de volta ao R

```
##### Querys adaptadas para o Spark
seg_dose_query <- glue("SELECT * FROM __THIS__
WHERE (vacina_descricao_dose='2ª Dose'
OR vacina_descricao_dose='2ª Dose Revacinação')")

fast_join <- glue("SELECT name_health_region as regioao_saude
FROM ({seg_dose_query}) as seg_dose
LEFT JOIN codigos
ON seg_dose.estabelecimento_municipio_codigo
= codigos.est_mun_codigo")

qnt_vax_query <- glue("SELECT regioao_saude, COUNT(*) AS N
FROM ({fast_join})
GROUP BY regioao_saude")
```

```

faixa_query <- glue("WITH qntVax AS ({qnt_vax_query})
                    SELECT regiao_saude, N,
                           CASE WHEN N > (
                                SELECT percentile_approx(N, 0.5)
                                FROM qntVax)
                           THEN 'Alto'
                           ELSE 'Baixo'
                           END AS Faixa
                    FROM qntVax")

bot5_query <- glue("WITH tabFaixa AS ({faixa_query})
                   SELECT regiao_saude, N, Faixa FROM
                   (
                       SELECT *, dense_rank()
                       OVER (PARTITION BY Faixa ORDER BY N DESC) as posicao
                       FROM tabFaixa
                   )
                   WHERE posicao <= 5")

##### Resultados em Memória
solve_spark <- function() {
  tbl_vax %>%
    ft_sql_transformer(bot5_query) %>%
    collect()
}
solve_spark()

```

e) Compare o tempo de processamento das 3 abordagens (SQLite, MongoDB e Spark), desde o envio do comando sql até o recebimento dos resultados no R. Comente os resultados incluindo na análise os resultados obtidos no item 2d) da Lista 1.

Cuidado: A performance pode ser completamente diferente em outros cenários (com outras operações, diferentes tamanhos de tabelas, entre outros aspectos).

Solução:

Tabela 1: Benchmark dos procedimentos utilizados nos itens b), c) e d).

Solução	Mínimo (s)	Média (s)	Máximo (s)
SQLite	294.7	308.2	317.6
Mongo	69.5	69.9	70.8
Spark	4.1	4.8	5.5

Tabela 2: Benchmark do tempo de execução das funções da Lista 1 para lidar com o banco de vacinados e regiões de saúde.

Pacote	Mínimo (s)	Média (s)	Máximo (s)
data.table	5.21	5.42	5.56
dtplyr	1.86	1.97	2.13

Comparando então com a solução da Lista 1, é possível observar que as soluções em **SQLite** e o **MongoDB** apresentadas na Lista 2 se saíram “mal” em tempo computacional, possivelmente por diferença no acesso ao hardware (o número de *cores* que o **SQLite** tem acesso é diferente do **Spark**) ou nas mecânicas utilizadas para performas os cálculos.

Por outro lado, o **Spark** chegou a desempenhar melhor que o **data.table** e pior que o **dtlyr**, valendo lembrar ainda que o **Spark** oferece possibilidade de configurar o número de *cores* disponíveis para o processamento, no caso foram utilizados 8 *cores*. E comparando os métodos da Lista 2 apenas entre eles, observa-se que o **Spark** teve o melhor desempenho disparado.

Repositório Contendo o Código Fonte

<https://github.com/Voz-bonita/Big-Data-Compute>