Banco de dados Desafio_03

Luiz Fernando de Oliveira Pereira RA: 267356

Quarto

Arquivo Parquet

- É um formato colunar para armazenar dados dados que foi criado em 2013 inicialmente pelo Twitter e Cloudera
- Seu principal objetivo é a leitura e compreensão de grande quantidade de dados (Big Data)
- Usado em diversas áreas como Spark, Hadoop, AWS, Azure, Google BigQuery, bancos analíticos e dentre outros
- Suas Facilidades consistem em
  - Ter um armanezamento mais compacto, sendo colunar e comprido;
  - Proporcionar uma leitura rápida de apenas algumas colunas sem a necessidade de carregar todo o dataset.

Pontos para se ter atenção é que por serem arquivos binários, não são diretamente legíveis e serem pouco útil para pequenos datasets.

Arquivo JSON (JavaScript Objet Notation)

- Seu formato é em texto, foi criado em 2001 por Douglas Crockfor.
- Seu objetivo é armazear e transmitir dados de forma estruturada e legível, proporcionando uma leitura e escrita simples.
- Pelo nome JSON, foi inspitado no JavaScript
- Usado em áreas como APIs web, configuração de sistemas, intercâmbio de dados entre linguages
- Sua facilidades consistem em

- Proporcionar uma leitura, edição e compreensão mais fácil;
- É aceito praticamente em todas as linguagens

Não é tão eficiente em compreesão quanto o parquet e a presença de estruturas aninhadas podem ser complexas para manipular em tabelas.

Instalando a biblioteca Arrow, utilizada para importar arquivos .parquet

```
#| echo: false
#| message: false
#| warning: false
install.packages("arrow", repos = "https://cloud.r-project.org/")
```

```
Installing package into 'C:/Users/ra267356/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
```

```
  There is a binary version available but the source version is later:
       binary   source needs_compilation
arrow 19.0.1.1 21.0.0.1              TRUE

  Binaries will be installed
package 'arrow' successfully unpacked and MD5 sums checked
```

```
Warning: cannot remove prior installation of package 'arrow'
```

```
Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
C:\Users\ra267356\AppData\Local\R\win-library\4.3\00LOCK\arrow\libs\x64\arrow.dll
to C:\Users\ra267356\AppData\Local\R\win-library\4.3\arrow\libs\x64\arrow.dll:
Permission denied
```

```
Warning: restored 'arrow'
```

```
The downloaded binary packages are in
    C:\Users\ra267356\AppData\Local\Temp\RtmpOUkg7Y\downloaded_packages
```

A biblioteca arrow para importar arquivos .parquet e a biblioteca jsonlite para importar arquivos .json

```
#| echo: false
#| message: false
#| warning: false
library(arrow)
```

```
Warning: package 'arrow' was built under R version 4.3.3
```

```
Attaching package: 'arrow'
```

```
The following object is masked from 'package:utils':

    timestamp
```

```
library(jsonlite)
```

Aqui um exemplo de como criar um data frame direto do R e deixá-lo no formato parquet

```
#Criando um data frame simples
df <- data.frame(
  id = 1:3,
  nome = c("Ana", "Bruno", "Carlos"),
  idade = c(23, 35, 29)
)

#Salvando em parquet
write_parquet(df, "exemplo.parquet")

# Lendo o arquivo parquet
dados_parquet <- read_parquet("exemplo.parquet")
head(dados_parquet)
```

```
  id   nome idade
1  1    Ana    23
2  2  Bruno    35
3  3 Carlos    29
```

Lendo o arquivo weather.parquet, que está no formato parquet, onde os dados tratam do histórico de medições meteorológicas de várias estações.

```r
# Lendo o arquivo parquet
parquet_weather <- read_parquet("weather.parquet")

head(parquet_weather)
```

```
  MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir WindGustSpeed
1     8.0    24.3      0.0         3.4      6.3          NW            30
2    14.0    26.9      3.6         4.4      9.7         ENE            39
3    13.7    23.4      3.6         5.8      3.3          NW            85
4    13.3    15.5     39.8         7.2      9.1          NW            54
5     7.6    16.1      2.8         5.6     10.6         SSE            50
6     6.2    16.9      0.0         5.8      8.2          SE            44
  WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm
1         SW         NW            6           20          68          29
2          E          W            4           17          80          36
3          N        NNE            6            6          82          69
4        WNW          W           30           24          62          56
5        SSE        ESE           20           28          68          49
6         SE          E           20           24          70          57
  Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm RainToday RISK_MM
1      1019.7      1015.0        7        7    14.4    23.6        No     3.6
2      1012.4      1008.4        5        3    17.5    25.7       Yes     3.6
3      1009.5      1007.2        8        7    15.4    20.2       Yes    39.8
4      1005.5      1007.0        2        7    13.5    14.1       Yes     2.8
5      1018.3      1018.5        7        7    11.1    15.4       Yes     0.0
6      1023.8      1021.7        7        5    10.9    14.8        No     0.2
  RainTomorrow
1          Yes
2          Yes
3          Yes
4          Yes
5           No
6           No
```

Lendo o arquivo no formato json chamado de employess.json que trata da coleta de dados relacionados a funcionários de uma empresa

```r
# Lendo o arquivo JSON
dados_jsonN <- fromJSON("employees.json")

head(dados_jsonN)
```

```
  employee.id       employee.name   employee.position employee.department.id
1     E00001         Traci Kelly Geologist, wellsite                    D006
2     E00002      Thomas Mendoza Magazine journalist                    D085
3     E00003    Natasha Williams     Tourism officer                    D083
4     E00004        Walter Hanna       Herpetologist                    D076
5     E00005     Daniel Williams   Social researcher                    D049
6     E00006 Mr. David Rodriguez    Fashion designer                    D076
  employee.department.name employee.department.manager.id
1                Monetize                           M1179
2                 Harness                           M6539
3                 Harness                           M7225
4                Expedite                           M5714
5               Cultivate                           M2305
6               Transform                           M9031
  employee.department.manager.name employee.department.manager.contact.email
1                    Alicia Hughes                    patricia05@example.net
2                      Bryan Jones                       leeeric@example.net
3                  Stacey Johnston            williamcunningham@example.net
4                      Mark Acosta                    zfletcher@example.com
5                   Misty Cummings                       erin74@example.net
6                    William Olsen                    dorothy11@example.com
  employee.department.manager.contact.phone
1                                2735306585
2                      (461)803-4754x44615
3                        (374)225-2216x990
4                            (698)802-9911
5                     001-713-617-8609x1921
6                        (646)685-9737x8653

1          P3213, Function-based zero-defect application, 2025-06-24, T292, T465, Disintermed
End Action-Items, Completed, Completed, 18, 17, Scikit-learn, Python, Pandas, TensorFlow, Sci
learn, SQL, Python, NumPy, 2025-08-18, 2025-07-12
2                                      P6704, Expanded logistical core, 2025-
07-08, T428, T860, Facilitate Revolutionary Markets, Enhance Turn-Key Markets, Pending, In Pr
learn, Python, JavaScript, NA, 2025-09-10
3                                        P4328, Synergistic actuating help-
desk, 2025-07-21, T428, T617, Transform Innovative Models, Scale Killer Channels, In Progress
learn, Pandas, NumPy, SQL, 2025-08-21, NA, NA, 2025-09-14
4                                                                P1496, Polarized execu
06-22, T990, T608, Redefine Extensible Platforms, Engage Turn-Key Partnerships, Pending, In P
09-07
5 P3965, Pre-emptive interactive pricing structure, 2025-08-10, T215, T620, Optimize Out-
Of-The-Box Supply-Chains, Facilitate User-Centric Networks, Completed, Completed, 44, 64, SQL
```

11-02, 2025-10-08
6 P8367, Optional asymmetric installation, 1
10-14, T687, T991, Redefine Dot-Com Infrastructures, Streamline Plug-And-
Play Vortals, Pending, Completed, 60, 71, Python, Java, TensorFlow, Pandas, PyTorch, Java, NA
01-02