

Servidores estatísticos

Servidores de alta capacidade com softwares para processamento de dados:

servidor	softwares
bsb_stat1	r, python, stata
bsb_stat2	r, stata, dbeaver, spss
bsb_stat3	r, stata, dbeaver
bsb_stat4	r, python
bsb_jupyterhubGPU	python
sasworkspace1	sas
rio_stat1	r, python, stata, dbeaver

*Nomes antigos dos servidores, na ordem: sstata1, sstata2, sstata3, bsb_stat, jupyterhubGPU, sasworkspace1, srjd0.

Como acessar? 1) Solicitar acesso por e-pedidos de TI; 2) Estar na rede-Ipea (PC ou conexaoVPN); 3) Acesso remoto ao servidor. O *jupyterhub* é acessado pelo browser.

Bases de dados

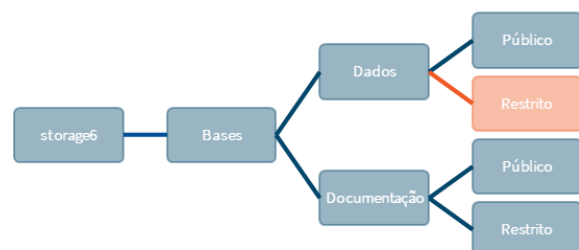
Quais são as bases disponíveis? Veja o catálogo na [Intranet](#) ou em nuvem.ipea.gov.br.

storage	localização*
storage6	bsb
storage1	bsb
psql10-df/rais	bsb
srjn4	rio

*Os *storages* devem ser acessados por meio de servidores do mesmo local (Brasília ou Rio).

Arquitetura das pastas de dados (storage6)

As versões originais das bases de dados ficam na pasta storage6/dados. Para dados sigilosos, explore a documentação mesmo antes de ter solicitado acesso ([Saiba como solicitar acesso a bases restritas](#)).



10 mandamentos:

1. USARÁS a memória RAM com parcimônia

Carregue apenas variáveis e observações essenciais. Limpe periodicamente o ambiente (`gc()` = garbage collector). Em análises com grandes volumes de dados, teste rotinas completas com pequenas amostras.

2. NÃO GUARDARÁS cópias de bases de dados em vão

Evite redundâncias, especialmente com bases grandes e de acesso restrito, como RAIS e CadÚnico.

3. NÃO SALVARÁS na área de trabalho dos servidores

A área de trabalho fica no disco “C:” que é compartilhado. Se encher, trava o servidor para todos. Utilize diretórios na rede e/ou repositórios de código para armazenar scripts, outputs e dados (estes só na rede).

4. NÃO USARÁS seu PC para processar e salvar dados

PCs não têm capacidade, backup, redundância elétrica ou isolamento físico. Use os servidores estatísticos.

5. NÃO SALVARÁS dado restrito em pasta compartilhada

Todos o usuário com acesso ao diretório onde salvar base identificada precisa ter permissão de acesso a esta base.

6. NÃO RETIRARÁS dados restritos da rede do Ipea

Utilize-os exclusivamente dentro de servidores na rede do Ipea, sem copiá-los para qualquer dispositivo.

7. USARÁS o servidor menos congestionado

Para identificar *heavy users* : **Task Manager>More Details>Users** e cole o *username* (R*, B* ou T*) na busca do *Webmail*.

8. USARÁS os servidores apenas para análise e modelagem de dados

Para internet, intranet, IpeaProjetos use seu PC ou desktop virtual. Não use para fins não institucionais (ex. treinar algoritmo de ML para trabalho de curso).

9. NÃO TRAFEGARÁS bases de dados entre o Rio e Brasília

Evite abrir dados armazenados em *storages* de Brasília por meio de um servidor do Rio de Janeiro e vice-versa.

10. PROCESSARÁS dados de forma eficiente

Para R, sugerimos **data.table**, **arrow**, **DuckDB**, ou bancos de dados (**SGBD-SQL**), conforme benchmark a seguir.






***O descumprimento dos mandamentos 2, 4 e 5 pode acarretar consequências legais, previstas na Lei Geral de Proteção de Dados (LGPD).**

Manipulação e modelagem: :pacotes recomendados

Benchmark:

Simulação com dados da RAIS vínculos 2004 (44 milhões de linhas): leitura, tabulação de empregados por setor e UF, e estimação de um modelo de regressão minceriano (ols, iv, e fe).

Leitura e tabulação por setor e UF
(44 milhões de observações)*

pacote	dados	Mínimo (min)	Mediana (min)
 arrow	parquet	1.04	1.08
 sgbd	psql	1.25	1.32
 data.table	csv	2.81	2.85
 duckdb**	csv/sql	2.98	3.08
 dplyr	csv	3.92	4.02

*10 iterações: leitura, tabulações e estimação de um modelo de regressão.

**No duckdb, os tempos consideram a criação, em disco, de uma versão SQL da base de dados. Depois dessa etapa, os processos de leitura e manipulação de dados são semelhantes ao arrow e ao sgbd.

Estimação de equação minceriana
(400 mil observações)

pacote	Padrão (s)	efeito fixo * (s)	IV (s)
fixest	1.24	1.26	4.02
lfe	2.81	4.07	6.03
lm (base r)	2.82	305	--

*Efeitos fixos para 561 CNAEs.

Os resultados completos estão no [GIT do IpeaDATA-lab](https://github.com/IpeaDATA-lab).

ipea

Instituto de Pesquisa
Econômica Aplicada



DuckDB

- > Sintaxe do *dplyr* ou SQL
- > Pode utilizar banco SQL local, criado a partir dos dados originais.

Saiba mais:

<https://duckdb.org/docs/api/r>

Ex. Computando vínculos formais da RAIS por CNAE:

```
#Leitura CSV com data.table
dados <- fread("path_rais_csv", select =
  "clas_cnae10")
```

```
#Escrever os dados em BD SQL local
con <- dbConnect(duckdb(), "path_db_sql")
dbWriteTable(con, "brasil2004", dados,
  overwrite = TRUE)
```

```
#Número de empregados por setor CNAE
tab_cnae <- tbl(con, "brasil2004") %>%
  count(clas_cnae10, name =
    "num_empregados") %>% collect()
```



- > Sintaxe do *dplyr*.
- > Não é necessário subir bases completas para a memória.

Saiba mais:

<https://arrow.apache.org/docs/r/>

Ex. Calculando o número de vínculos a RAIS por UF:

```
#Leitura dos dados "fora da memória"
dados <- open_dataset("path_rais_parquet")
```

```
#Filtrar colunas relevantes
dados <- dados %>%
  select("uf")
```

```
#Número de empregados por UF
tab_uf <- dados %>%
  count(uf, name = "num_empregados")
```

```
#Retornar resultado
tab_uf <- tab_uf %>% collect()
```

fixest

- > Estima modelos lineares e glm
- > Suporta qualquer número de efeitos fixos

Saiba mais:

<https://lrberge.github.io/fixest/>

```
#Setup
dados <- open_dataset("path_rais_parquet")
dados <- dados %>% select("clas_cnae10",
  "género", "raca_cor",
  "grau_instr",
  "rem_med_r",
  "temp_empr") %>%
  collect()
```

```
#Rodar modelo com efeitos fixos (FE)
#e variável instrumental (IV)
feols(rem_med_r ~ genero + temp_empr +
  I(temp_empr^2) | clas_cnae10 |
  grau_instr ~ raca_cor,
  data = dados)
```

FE

IV