




FRAMEWORK PARA PROCESSAMENTO DE DADOS – MONITORAMENTO DE DOENÇA INFECCIOSA

3

*Paula Eduarda de Lima
Mariana Fernandes Rocha
Ana Júlia Amaro Pereira Rocha
Maria Eduarda Mesquita Magalhães*



OBJETIVO

Criar um micro-framework para desenvolver pipelines de processamento de dados concorrentes e paralelos a fim de monitorar doenças infecciosas.

01 Arquitetura do micro-framework

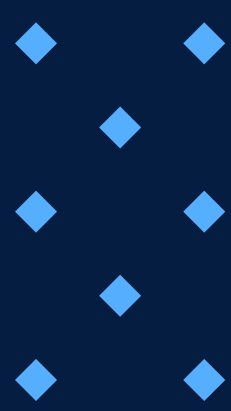
02 Observações sobre o projeto

03 Fluxograma

04 Sobre os dados

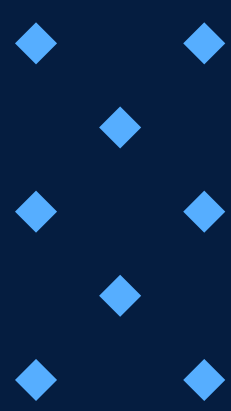


ARQUITETURA DO MICRO-FRAMEWORK

- **Mock** → Nossas fontes de dados serão simulações de hospitais, da Secretaria de Saúde (SS) e da Organização Mundial da Saúde (OMS). Receberemos um conjunto de dados advindos de pelo menos uma dessas entidades toda semana.
 - **Triggers** → A ideia inicial é que esta seja uma classe com os atributos Timer e Request representando os dois tipos de triggers que teremos. Além disso, para evitar condição de corrida, usaremos a lógica do produtor-consumidor ao transferir arquivos de um passo a outro no pipeline. Por exemplo, no momento em que os arquivos saem do extrator e destinam-se ao primeiro tratador geral.
- 

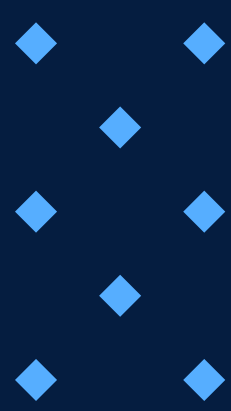


ARQUITETURA DO MICRO-FRAMEWORK

- **Extractor** → Haverá três extratores, um para cada tipo de arquivo (TXT, CSV e SQLite). A etapa de extração dos dados converte os arquivos de diferentes tipos em DataFrames, todos no mesmo padrão.
 - **Tratador** → Definimos 6 tratadores, sendo 3 gerais, que seriam aplicados em todos os DataFrames e 3 específicos, para tratamentos mais individuais de alguns dados da saúde. Todos os arquivos passam pelos 3 tratadores gerais e podem passar por um ou mais dos outros tratadores específicos ou ir direto para o loader.
- 

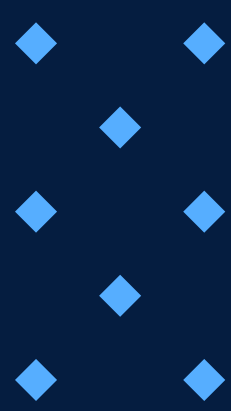


OS TRATADORES

- **Tratadores gerais:**
 - **T1** → Filtra valores inválidos, isto é, idade negativa por exemplo;
 - **T2** → Remove linhas majoritariamente nulas;
 - **T3** → Remove colunas majoritariamente nulas.
 - **Tratadores específicos:**
 - **T1** → Faz merge entre dois datasets por uma coluna em comum;
 - **T2** → Agrupa um dataset dado uma coluna de parâmetro (ex: por região);
 - **T3** → Calcula a média dado uma coluna de parâmetro e cria uma coluna bool que é verdadeira quando uma observação está acima da média. (ex: hospitais acima da média do número de infectados da região)
 - **OBS:** Os arquivos podem passar por qualquer desses tratadores, mas se passarem pelo T1 não passam pelo T2 e vice-versa.
- 

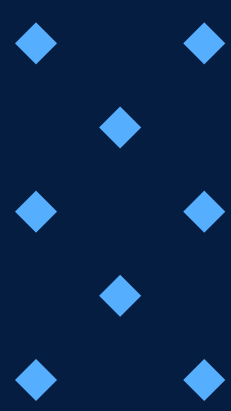


LOADER E DISPLAY

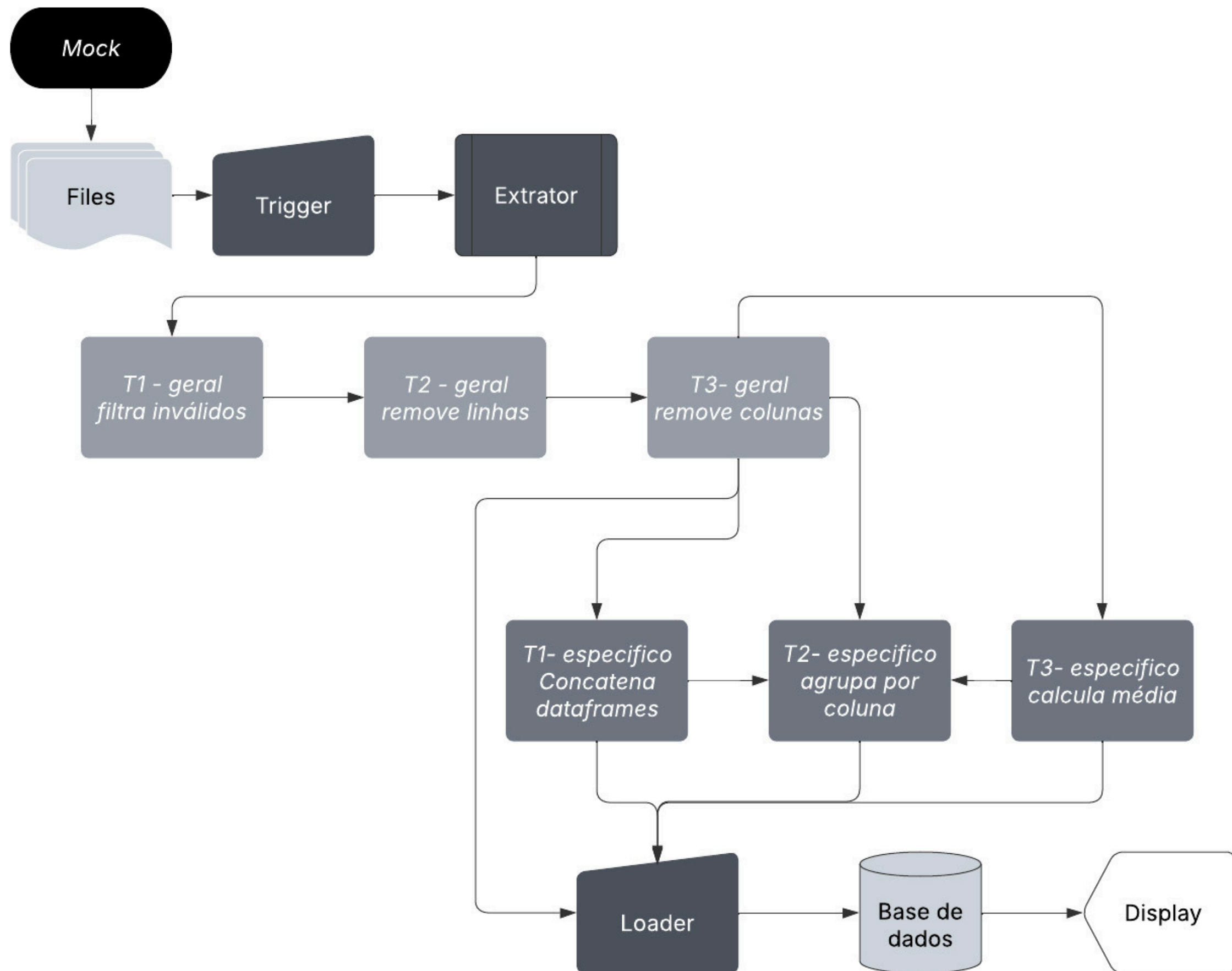
- **Loader** → Carrega os arquivos processados para o banco de dados de modo que agora possam ser acessados e analisados confiavelmente.
 - **Display** → A partir do Banco de Dados faremos algumas análises simples, como mostrar hospitais que estão com o número de infectados acima da média da região e mostrar a soma de casos por região. Essas análises aparecerão em um dashboard ou mesmo no console e são para teste do uso dos dados . Além disso, também serão impressos os tempos de execução para visualizar a eficiência do nosso modelo.
- 



OBSERVAÇÕES SOBRE O PROJETO

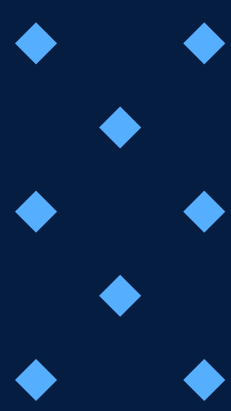
- Extratores e tratadores farão uso de paralelismo;
 - Ao transferir arquivos de um passo a outro no pipeline sempre usaremos a lógica do produtor–consumidor para maior eficiência do modelo;
 - O projeto será implementado em C++;
 - As simulações de dados serão feitas em python.
- 

FLUXOGRAMA





SOBRE OS DADOS – OMS

- Nossos dados serão referentes a algumas ilhas as quais são subdivididas em regiões.
 - Todas as ilhas possuem hospitais mas nem todas as regiões possuem.
 - A OMS vai nos fornecer dados gerais, ou seja, por ilha, sendo que as variáveis em seus datasets serão:
 - Data
 - Número de óbitos
 - População
 - CEP da ilha
 - Número de recuperações
 - Número de vacinados
- 

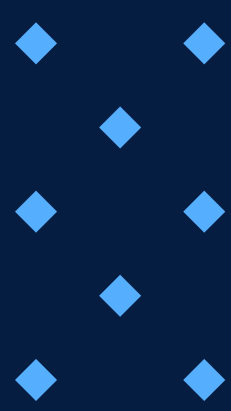


OMS

Data (dia-mes-ano)	Nº óbitos (int)	População (int)	CEP ilha (int)	Nº recuperados (int)	Nº de vacinados (int)



SOBRE OS DADOS – SS

- A Secretaria de Saúde (SS) fornecerá dados por região de cada ilha, sendo que as variáveis presentes em seus datasets serão:
 - Data
 - Vacinado (sim/não)
 - Diagnóstico (positivo/negativo)
 - Nível de escolaridade do paciente (por um código numérico)
 - População regional
 - CEP (ilha+região)
 - A ideia é que a Secretaria nos apresente mais sobre os pacientes que fizeram teste para a doença, informando se este já foi vacinado, entre outras informações.
- 

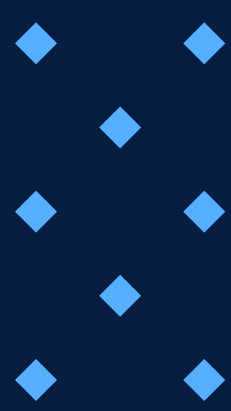


SECRETARIA DA SAÚDE

Data (dia-mes-ano)	Diagnóstico (bool)	Vacinado (bool)	CEP (int)	Escolaridade (int)	População (int)



SOBRE OS DADOS – HOSPITAIS

- Os hospitais fornecerão dados mais individuais e locais dentro de cada região, sendo que as features presentes em seus datasets serão:
 - Data
 - Internado (sim/não)
 - Idade
 - Sexo (1 para FEM e 0 para MAS)
 - CEP (ilha+região)
 - sintoma 1 (teve/não teve)
 - sintoma 2 (teve/não teve)
 - sintoma 3 (teve/não teve)
 - sintoma 4 (teve/não teve)
- 



HOSPITAIS

Data(dia-mes-ano)	Internado (bool)	Idade (int)	Sexo (int)	CEP (int)	Sintoma1 (int)	Sintoma2 (int)	Sintoma3 (int)	Sintoma4 (int)



PROTÓTIPO EM CÓDIGO

Na apresentação de segunda-feira, iremos implementar uma estrutura de pipeline para processar os dados de monitoramento da doença infecciosa.

Uma aplicação testada será em um DataFrame resultante que identificará as regiões onde o número de casos registrados ultrapassa a média geral de todo o território monitorado.

E a outra análise plausível é de regiões mais infectadas em um certo período de tempo.

