Desenvolvimento de um Modelo de Ingestão de Dados para AutoML

Guilherme Albuquerque[1]
orcid.org/0000-0003-0610-5834

Gabriel Mac'Hamilton[1]
orcid.org/0000-0002-3735-190X

Alexandre Maciel[1]
orcid.org/0000-0003-4348-9291

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: gtra@ecomp.poli.br

DOI: 10.25286/repa.v7i3.2457

Como citar este artigo pela NBR 6023/2018: Guilherme Albuquerque; Gabriel Mac'Hamilton; Alexandre Maciel. Desenvolvimento de um Modelo de Ingestão de Dados para AutoML.Revista de Engenharia e Pesquisa Aplicada, Recife, v. 7, n. 3, p. 29-38.

RESUMO

A interação entre alunos e professores em Ambientes Virtuais de Aprendizagem (AVA), produzem dados do tipo educacional, que possuem alto potencial de análise, contudo, a extração dessas informações para ferramentas de estudo como as de Aprendizagem de Máquina Automatizada (AutoML), demanda conhecimentos técnicos avançados, que não encontramos em usuários comuns, sendo assim, este trabalho busca adicionar uma nova forma de ingestão das informações em uma ferramenta de AutoML, almejando diminuir a complexidade de um processo de Extração, Transformação e Carga (ETL) dos dados, utilizando a ferramenta *Pentaho Data Integration* (PDI).

PALAVRAS-CHAVE: AutoML; Ingestão de dados; ETL; PDI; AVA;

ABSTACT

The interaction with Vitual Learning Envinoments (AVA) made by teachers and students, produce data of educational kind, that posses an high potential of analysis, however, the extraction of these information to study tools like Automated Machine Learning (AutoML), demands advanced technical knowledge, which we do not find in common users, therefore, this works seeks to add a new way of ingestion of the informations in a AutoML tool, aiming to reduce the complexity of a process of data Extraction, Transformation and Loading (ETL), using the Pentaho Data Integration (PDI) software.

KEY-WORDS: AutoML; Data ingestion; ETL; PDI; AVA;

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Para continuar ativo na sociedade um indivíduo precisa estar apto e atualizado. Essa demanda educativa é similar a dos alunos frequentadores de escolas e universidade **[1].**

Para suprir essa demanda, era necessário recorrer aos meios tradicionais, como aulas presenciais e livros físicos, mas com as facilidades do avanço da tecnologia surgiram novas possibilidades que facilitaram o ensino, como vídeo aulas, livros em formato digital, entre outros. Em particular, temos os Ambientes Virtuais de Aprendizagem (AVAs), que fornecem uma experiência simulada de uma sala de aula tradicional. Em um AVA os participantes se encontram em tempo real para assistir as aulas; possuem um espaço para tirar dúvidas; outro para entrega de tarefas; inclusive facilidades que não são encontradas na sala de aula tradicional como, por exemplo, assistir uma aula que foi gravada, facilitando assim a fixação do conhecimento.

Atualmente vivemos em um mundo onde os dados são considerados mais preciosos que o petróleo **[2]** e a exploração dos dados gerados por ferramentas de AVAs tornou-se objeto de estudo, possibilitando *insights* e previsões provenientes dessas fontes. Um conglomerado de estudos focados em ferramentas de ensino online, promoveu a construção de um *framework* de Aprendizagem de Máquina Automatizado (AutoML) chamado *Framework* de Mineração de Dados Educacionais (FMDEV), que visa a aplicação de técnicas de mineração de dados e *machine learning* de uma forma que seja simples para o usuário final sem ser necessário o conhecimento técnico para desenvolvimento de código de aprendizagem de máquina **[3]**.

A ingestão dos dados provenientes dos AVAs para o *framework*, necessita de atenção, pois um usuário que não detenha conhecimentos técnicos na área de programação deve encontrar dificuldades para extrair os dados, realizar alguma transformação e a posterior carga no destino. Este processo de Extração, Transformação e Carga (ETL) ou sua variante ELT que inverte as etapas de transformação e carga serve para proporcionar o transporte de informações entre os bancos de dados distintos **[4]** sendo este o processo utilizado neste trabalho para a ingestão dos dados em um AutoML.

1.2 OBJETIVOS

Este estudo possui o seguinte objetivo: Criar e aplicar um modelo de ingestão de dados educacionais que possui como fonte os Sistemas de Gestão de Aprendizagem (LMS), e como destino uma ferramenta de AutoML. Para atingir este objetivo geral, serão buscados os seguintes objetivos específicos: construir um fluxo de ETL a partir da ferramenta *Pentaho Data Integration* (PDI), criar a nova funcionalidade de ingestão com os fenômenos de estudo na ferramenta de AutoML FMDEV e realizar testes de integração da solução desenvolvida.

1.3 JUSTIFICATIVA

As ferramentas de AutoML melhoram a produtividade na análise dos dados, oferecem uma blindagem contra erros metodológicos e democratizam o acesso a análise para profissionais que não são especialistas em aprendizagem de máquina **[5]**, sendo assim, faz-se necessário realizar um processo de ETL que busca facilitar o uso de uma ferramenta de AutoML.

De acordo com Hutter **[6],** a utilização de uma ferramenta de AutoML economiza uma grande quantidade de tempo e dinheiro, visto que um profissional de aprendizagem de máquina é caro e escasso. Desta forma, podemos concluir que este trabalho é importante por trazer retorno financeiro e redução de tempo.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 AVAS E ENGAJAMENTO ESTUDANTIL

A utilização de AVAs apresenta como efeito negativo o distanciamento entre o aluno e o professor, que possui dificuldades em identificar os seus alunos **[7].** Este cenário dificulta o reconhecimento de como os alunos estão em relação ao curso, mais precisamente como está o seu nível de engajamento.

Para sanar este problema, foi realizado um trabalho por Macêdo **[7],** no qual a utilização de técnicas de mineração de dados possibilitou a descoberta de perfis de engajamento, proporcionando assim, a medição da participação de cada aluno.

Dentre os tipos de engajamento existentes o que será referido neste estudo é o comportamental, que se refere às atitudes positivas do estudante em

relação ao cumprimento das regras e adesão às normas da sala de aula **[8].**

2.2 AUTOML

Dentro do estudo de Aprendizagem de máquina, possuímos um sub campo que vem se tornando cada vez mais independente, que é o AutoML. O principal objetivo do AutoML é automatizar completamente as tarefas de mineração de dados, aprendizagem de máquina, reconhecimento de padrões e análise de dados avançadas **[9]**. Este tipo de aplicação, caracteriza-se por ser um programa no qual os procedimentos realizados por ele dispensam o usuário de realizar suas atividades com linhas de código, para utilizar uma interface que recebe os parâmetros passados e aplica de forma automatizada as técnicas já existentes na ferramenta. Assim, uma ferramenta de AutoML necessita que os dados provenientes de algum sistema original sejam carregados na mesma. Essa ingestão de dados para uma ferramenta AutoML faz parte do processo de Pipeline de Dados e pode ser feita de diversas formas como ETL ou ELT.

2.3 ETL OU ELT

Para realizar a criação e execução de um fluxo de dados, necessitamos possuir a capacidade de realizar três ações distintas, que quando combinadas, se completam, possibilitando que a transferência das informações venha a ocorrer de forma satisfatória.

O termo ETL do inglês: *Extract Transform and Load*, faz menção a essas três etapas na seguinte ordem:

- Extrair, na qual realizamos a tarefa de busca e extração de uma ou mais bases, sendo elas as mais variadas possíveis, como banco de dados, arquivos csv, bases de dados em ferramentas de Sistemas Integrados de Gestão Empresarial, entre outros.
- Transformar, esta é a fase na qual aplicamos as transformações desejadas nos dados que buscamos na etapa anterior, estas transformações podem ser uma junção dos dados das diferentes fontes, uma exclusão de um campo visto que a mesma informação está presente em uma fonte diferente, também pode ser o tratamento de campos nulos ou até mesmo a troca do tipo de campo, tudo isto é feito para atender a necessidade da base destino da ingestão dos dados.
- Carregar, nesta etapa, ocorre a carga no local de destino definido na arquitetura de solução da atividade, este carregamento ocorre de acordo com as transformações realizadas na etapa anterior **[10]**.

Além do já citado ETL, temos outro modelo conhecido como ELT, no qual ocorre a inversão da ordem de execução das atividades executando primeiramente o processo de carga antes da transformação. Esta mudança possibilita que a transformação passe a ser realizada no repositório final, podendo ser mantida a forma original trazida da fonte para fins de consulta e duplicação dos dados, caso seja de interesse do usuário do sistema realizar estas atividades **[11]**.

2.3.1 Ferramentas de ETL/ELT

Os três passos do ETL são consolidados por meio de ferramentas que facilitam o processo, existindo a possibilidade de realizar as atividades em um mesmo software ou por meio de linguagens de programação, no qual é necessário um conhecimento em desenvolvimento de código mais avançado.

Os três passos do ETL são consolidados por meio de ferramentas que facilitam o processo, existindo a possibilidade de realizar as atividades em um mesmo software ou por meio de linguagens de programação, no qual é necessário um conhecimento em desenvolvimento de código mais avançado.

Uma das opções para construir um processo é quando o desenvolvedor elabora em linhas de código as rotinas de execução do ETL, geralmente utilizando scripts em diferentes linguagens. Por ser necessária a criação de código fonte, a utilização desse método é indicada para tarefas pequenas que não necessitem de controle para fluxos complexos, pois o tempo de desenvolvimento torna-se volumoso e não condizente com a expectativa de entrega por parte do cliente que contratou a solução, além disso, estes scripts costumam não utilizar de paralelismo, acarretando em uma execução demorada **[12]**. Para contornar estas dificuldades é recomendada a utilização das ferramentas, como por exemplo o Talend que é um *software* robusto que realiza todas as operações de ETL.

Outra opção que possui código aberto é o PDI que foi escrito em Java e é independente de plataforma, ou seja, ele funciona no Windows, Mac, Ubuntu,

DOI: 10.25286/repa.v7i3.2457

entre outros. Ele possui duas versões, a Community e Enterprise, sendo a última uma versão paga com suporte especializado **[13]**.

No PDI temos três partes principais, o Spoon que é a interface gráfica onde o usuário realiza a montagem do fluxo da sua solução, temos também o Pan, que é um motor de transformação de dados que performa operações de leitura, manipulação e escrita para e a partir de várias fontes de dados. A terceira parte principal é o Kitchen, que executa *Jobs* que são desenhados na ferramenta Spoon ou de um repositório com itens já inclusos **[13]**.

2.4 TRABALHOS RELACIONADOS

No trabalho de Zhou **[14]** é criado um AutoML chamado PAIRS AutoGEO, que é um framework voltado para dados geoespaciais. Nele, os dados são coletados pela ferramenta PAIRS, que realiza a agregação de diferentes fontes de dados. Para realizar a ingestão dos dados para o processo de aprendizagem de máquina, o usuário deve utilizar um arquivo no formato *JavaScript Object Notation* (JSON).

O arquivo JSON necessita de três informações para poder realizar a ingestão dos dados, a primeira é o tipo de dado que deseja ser obtido, a segunda são as coordenadas geográficas de latitude e longitude e a terceira é o intervalo de tempo da amostra **[14].**

No estudo de **[15]**, ocorreu a criação de um AutoML chamado Cardea, que possui como foco de processamento registros eletrônicos de saúde. A ingestão de dados neste AutoML é realizada por meio de linha de código, onde o usuário indica o caminho da pasta que contém o arquivo com os dados a serem ingeridos na ferramenta.

3 MATERIAIS E MÉTODOS

3.1 FMDEV

No trabalho de Silva **[16],** ocorre o desenvolvimento do FMDEV, que é um ambiente de Aprendizado de Máquina Automatizado e integrável a múltiplos Ambientes Virtuais de Aprendizagem **[16].**

Este *framework* foi utilizado em nosso estudo como a ferramenta de AutoML que ingeriu os dados a partir do nosso fluxo de ETL.

3.2 BASE DE DADOS

A base de dados utilizada neste projeto provém de estudantes de quatro cursos diferentes do Núcleo de Ensino a Distância da UPE (NEAD), estas informações são a fonte do nosso processo de ETL e o destino é o FMDEV.

Os dados do NEAD estão disponíveis em uma instância local de MySQL que é o banco de dados utilizado como solução da ferramenta LMS Moodle **[17]**.

3.3 AMBIENTE DE DESENVOLVIMENTO

Dentro do FMDEV, empregamos o PostgreSQL como solução para o banco de dados, Python para *back-end* e *React* para *o front-end*.

A troca de informações do *front-end* e o *back-end*, foi escrita utilizando a forma de organização de arquivos conhecida como *Ducks Modular Redux*. **[18]**.

A versão do PDI utilizada neste estudo foi a 9.2, este *software* não requer instalação e ao realizar o *download* dos arquivos, temos a versão para Windows e Linux disponibilizadas. Para o desenvolvimento do código neste estudo, foi utilizado o Subsistema Windows para Linux versão 2 com a imagem do Ubuntu 18.04.

Para executar os comandos nos bancos de dados em nosso projeto, fizemos uso do terminal e do programa DBeaver, que é um aplicativo de *software* no qual nos conectamos aos bancos de dados de diferentes distribuições.

3.4 METODOLOGIA

Para realizar a metodologia deste trabalho, utilizamos das seguintes etapas presentes no diagrama elaborado na Figura 1.

Figura 1 – Diagrama da metodologia

Fonte: Os autores.

Primeiro passo: O início do desenvolvimento ocorreu com o *download* do FMDEV presente no repositório oficial **[3]** e a instalação de acordo com o passo a passo disponível.

Segundo passo: Para receber a base NEAD, utilizamos de um banco de dados MySQL que foi configurado segundo instruções oficiais da Microsoft **[19]**. Após as configurações, os dados do NEAD foram importados do repositório do Grupo de Pesquisa em Ciência de Dados e *Analytics* (GPCDA) para o banco.

Terceiro passo: Para adicionar a aba de ingestão e seu fluxo na ferramenta do FMDEV, foram realizadas alterações nas linhas de código a começar pelo *front-end* no arquivo *index.js* da pasta *DataSource*, nele, criamos os modelos dos cartuchos relativos aos fenômenos educacionais e definimos a lógica de processamento quando esta opção for escolhida.

Operamos no banco de dados PostgreSQL a criação da tabela *controle_etl*.

No *back-end,* manuseamos o arquivo *Model.py* para adicionar a estrutura da tabela de controle, *controle_etl,* que define os cartuchos dos fenômenos educacionais. Esta estrutura será utilizada pelo arquivo *Etl.py* para buscar no banco os as informações e devolver ao *front-end* para serem apresentadas ao usuário.

Operamos no banco de dados PostgreSQL a criação da tabela de *indicators_engajamento*.

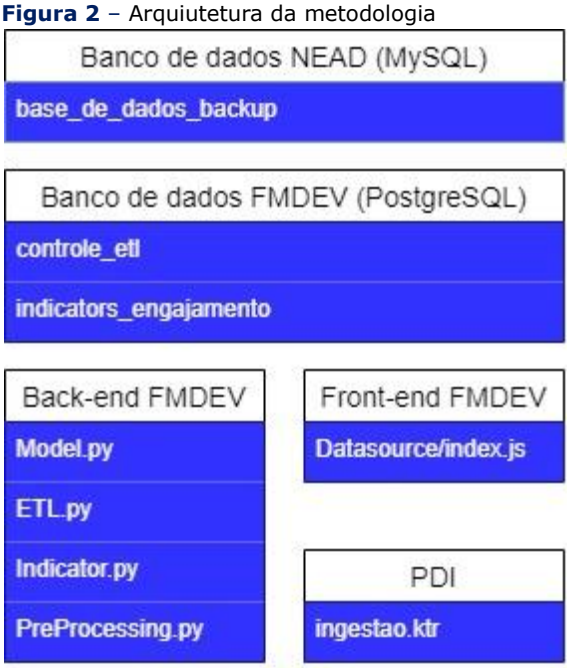
Para seguir com o fluxo, manipulamos o arquivo *Indicator.py* presente no *back*-end para adicionar a funcionalidade de buscar do banco de dados os campos e descrições referentes ao engajamento, que são apresentados na tela de seleção de indicadores.

Quarto passo: Utilização da interface gráfica disponível pelo Spoon para criação do fluxo de ETL, aplicando as seguintes conexões para o MySQL, host:127.0.0.1; nome do *Database*: moodle_backup; porta: 3306 e para o PostgreSQL, host:127.0.0.1; nome do *Database*: fmdev; porta: 5432.

Quinto passo: Manuseamos no *back-end* o arquivo *PreProcessing.py* que passou a possuir mais uma função, que é a de chamar a execução do processo de ETL criado no PDI.

O desenho da arquitetura da metodologia desenvolvida está de acordo com os itens criados ou modificados representados na figura 2.

Figura 2 – Arquiutetura da metodologia



Fonte: Os autores.

4 RESULTADOS E DISCUSSÕES

4.1 MODELO DE INGESTÃO

Para realizar o fluxo de dados, foi definido que seria realizada a extração, transformação e carregamento das informações relacionadas ao perfil de engajamento definido por Macêdo **[7]**, para isto, foi criada uma nova aba na área de Fontes de Dados dentro do FMDEV, com o nome de

DOI: 10.25286/repa.v7i3.2457

ingestão de dados, na qual conseguimos realizar o nosso trabalho de ingestão dos perfis de engajamento.

Figura 3 – Estado do banco de dados para tabela de controle

```
SELECT id, phenomenon, description
FROM public.controle_etl;
```

id	phenomenon	description
1	estudo	vem do moodle
0	engajamento	variaveis provindas do moodle

Fonte: Os autores.

Possuímos no banco de dados do FMDEV a tabela de controle definida por controle_etl conforme mostra a Figura 3 e a partir dela é realizada a leitura das informações para exibição na tela de Fontes de Dados apresentada na Figura 4.

Ao escolher o quadro de engajamento apertando no símbolo de *play*, teremos a próxima tela para escolha dos indicadores que são mapeados pela tabela de indicadores de engajamento conforme mostrado na Figura 5.

Figura 5 – Tabela de definição dos indicadores

```
SELECT id, "name", description, etl, created_at, updated_at
FROM public.indicators_engajamento;
```

id	name	description	etl
1	id_do_aluno	Identificação do aluno	engajamento
2	id_da_disciplina	Identificação da disciplina	engajamento
3	qtd_visu	quantidade de visualizações	engajamento
4	acessos_plataforma	acessos a plataforma	engajamento
5	entregues_dentro_prazo	atividades entregues dentro prazo	engajamento
6	postagens_forum	quantidade de postagens realizadas no forum	engajamento
7	desempenho	desempenho	engajamento

Fonte: Os autores.

A tela de escolha dos indicadores com informações provenientes da tabela mostrada na figura 5, pode ser vista na figura 6.

O perfil de engajamento possui os seguintes campos definidos: "Desempenho", "Quantidade de acessos à plataforma", "Atividades entregues no prazo", "Quantidade de postagens no fórum" e "Quantidade de visualização ao fórum". Porém os campos não estão escritos desta forma na base de dados, sendo necessário utilizar o dicionário de dados apresentado no Quadro 1.

Quadro 1 - Dicionário de dados

FONTE MOODLE	NOMENCLATURA VOLTADA AO USUÁRIO
DESEMPENHO	Desempenho
Var31	Quantidade de acessos à plataforma
Var33	Atividades engregues no prazo
Var34	Quantidade de postagens no fórum.
Var18	Quantidade de visualização ao fórum

Fonte: Repositório GPCDA.

Os atributos de identificação da disciplina e identificação do aluno foram adicionados aos outros cinco campos definidos, para obtermos uma forma de identificação do registro.

Como regra de negócio para o desenvolvimento da solução, temos que a tabela destino deve estar sempre atualizada para que as atividades realizadas estejam em dias sem dados defasados.

O desenho da solução compreende o manuseio do banco de dados local MySQL representando a utilização do banco do Moodle, que para uso em tempo real, é necessário apenas a troca das variáveis de conexão ao banco como endereço e porta, para que a conexão ao Moodle ocorra.

Para a realização do ETL, foi utilizada a ferramenta PDI. Dentro dela, fizemos uso dos seguintes componentes:

Execute SQL script: primeiro componente de todos utilizado no fluxo, para que ocorra a limpeza da tabela destino e os dados não tenham duplicidade. Por se tratar de uma operação realizada na base final, ele utiliza a conexão com o Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL.

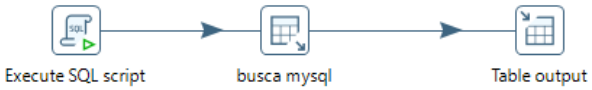
Table input: neste componente escrevemos o que desejamos extrair da fonte utilizando um comando de consulta, sendo ele: "select `ID do Aluno` as id_do_aluno, `ID da Disciplina` as id_da_disciplina, var18 as qtd_visu, var31 as acessos_plataforma , var33 entregues_dentro_prazo, var34 postagens_forum, cast(REPLACE(DESEMPENHO, ',','.') as decimal(4,3)) as desempenho from moodle_backup.base_de_dados_backup_csv".

Pode ser observado que nesta fase, já ocorreu a utilização do conceito da transformação, pois o campo "desempenho" estava em uma colunada definida como texto e no formato de definição de casas decimais do padrão de vírgula para casas de

milhar e ponto para decimal, sendo necessária a troca para vírgula como separador de casas de decimais e o ponto para casas de milhar, além disso, foi necessária a definição do tipo do campo como numérico, que para o banco de dados significa um campo numérico com ponto flutuante. A conexão utilizada para esta caixa foi a do MySQL, visto que estamos realizando a extração da fonte.

Table output: neste passo finalizamos nosso fluxo de dados realizando a carga na tabela definida dentro do componente com os dados que estão sendo transmitidos da fase anterior. Utilizamos a conexão com o PostgreSQL por ser o destino.

Figura 7 – Processo de ETL no PDI



Fonte: Os autores.

Este processo de ETL é chamado a partir *do back-end* pelo terminal ao ser selecionado o botão de pré-processar a base na tela de escolha de indicadores, permitindo que os dados sempre estejam atualizados e tudo isto ocorrendo sem requerer a abertura da interface do *Pentaho Data Integration* por meio do programa Spoon, dispensando o conhecimento técnico do usuário em desenvolvimento no PDI.

Ao final da execução, obtemos a carga completa na tabela de engajamento com o campo desempenho no formato designado, como podemos ver na Figura 8:

Figura 8 – Base de dados carregada no destino

```
SELECT id, id_do_aluno, id_da_disciplina, qtd_visu, acessos_plataforma
FROM public.engajamento;
```

amento 1 ×

T id, id_do_aluno, id_da_disciplina, qtd | Enter a SQL expression to filter results (use Ctrl+Sp

123 id_da_disciplina	123 qtd_visu	123 acessos_plataforma	123 entregues_de
116	15	133	
36	0	4	
49	8	48	
63	21	324	
30	27	151	
48	61	156	
31	8	43	
42	0	31	
9	38	246	

Fonte: Os autores.

Após este passo, a parte de ingestão de dados está completa e o programa segue o fluxo normal já preexistente.

4.2 TESTES DE INTEGRAÇÃO

Após fazer o login no FMDEV iremos até a área de Fontes de dados no símbolo de *mais* presente na aba da esquerda da ferramenta. Ao chegar nela e clicando em *Ingestão de dados,* veremos a aba que foi criada neste trabalho na figura 9.

Ao clicar no botão *play* iremos para a próxima tela apresentada na figura 10.

Nesta tela de indicadores temos a opção de voltar para a tela anterior ou escolher indicadores disponíveis e o indicador alvo para seguir com o fluxo do processamento da ferramenta, pressionando o botão *PRÉ PROCESSAR BASE* vamos para a seguinte tela, como podemos ver na figura 11.

A tela de Pré-processamento está aguardando o término do processo de ETL. Ao finalizar obtemos a tela que pode ser visualizada na figura 12.

A partir deste momento temos a opção de voltar para a tela anterior, ou seguir o fluxo previamente existente, dispensando a necessidade de mais testes.

4.3 RESULTADOS

Obtivemos como resultado do nosso estudo, uma nova forma de ingerir dados no *framework* por intermédio de uma aba desenvolvida no *front-end* e de um processo de ETL construído em cima do *software* PDI, a partir dele, conseguimos realizar o fluxo da ingestão de dados de forma simples e direta.

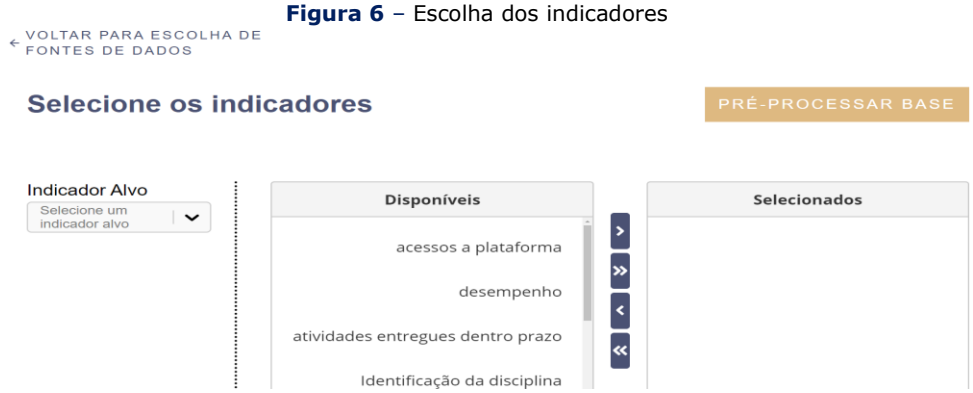
O transporte dos dados da fonte para o destino ocorreu em tempo hábil, levando cerca de 24 segundos para que o processo do Pan comece e mais 3 segundos para que o processo de carga seja realizado, consumando o transporte com mais de 30 mil linhas. Este resultado obtido foi bom pois o programa PDI utiliza de paralelismo, algo que geralmente não é encontrado em processos de ETL com o fluxo definido por diferentes *scripts* em linhas de código.

DOI: 10.25286/repa.v7i3.2457

Figura 4 – Tela inicial de escolha dos fenômenos

Fonte: Os autores.

Figura 6 – Escolha dos indicadores



Fonte: Os autores.

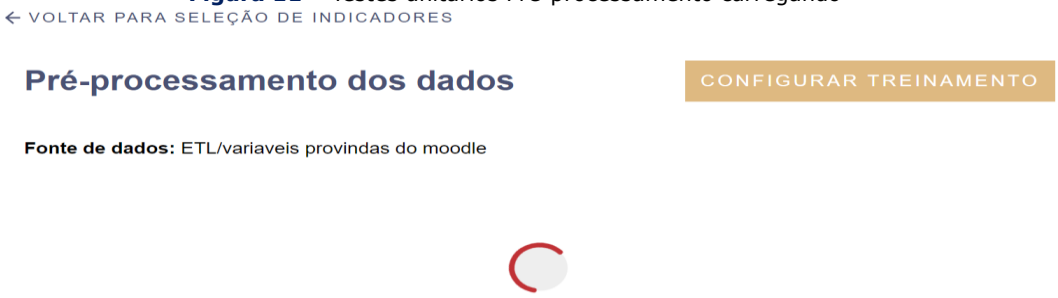
Figura 9 – Testes unitários fontes de dados

Fonte: Os autores.

Figura 10 – Testes unitários indicadores

Fonte: Os autores.

Figura 11 – Testes unitários Pré-processamento carregando



Fonte: Os autores.

Figura 12 – Testes unitários Pré-processamento finalizado

Fonte: Os autores.

4.4 DISCUSSÕES

Foi possível simplificar e diminuir o nível técnico necessário para o usuário final utilizar o FMDEV, pois, não se faz mais necessário, a extração manual realizando uma consulta no ambiente do Moodle, seguida de uma exportação dos dados e posterior importação manual no FMDEV para estudo do fenômeno de engajamento. Desta forma, este trabalho alcança os objetivos definidos com os resultados obtidos.

Também vale salientar que o fluxo desenvolvido apesar de ser simples, com apenas três componentes, ele é direto e eficaz, diminuindo o tempo de desenvolvimento da solução, assim como mitigando a possibilidade de erros devido a um desenho de solução complexa e desnecessária.

5 CONCLUSÕES e TRABALHOS FUTUROS

5.1 CONCLUSÕES

Dado o exposto que a extração dos dados de suas fontes e ingestão deles no destino final é uma tarefa que requer conhecimentos específicos que o usuário comum não possui, fica clara a necessidade da utilização de ferramentas de ETL para este propósito, sendo fundamental em tarefas que são repetitivas e que podem ser automatizadas, visando o ganho de tempo que temos para realizar o transporte da informação quando possuímos o fluxo dos dados já construído. Por isso, concluímos que a nova aba de ingestão desenvolvida com o fluxo de dados funcionando sem a necessidade de conhecimento técnico do usuário em ingestão de dados, democratiza a utilização do FMDEV para usuários que não possuem requisitos técnicos avançados em manuseio de dados.

5.2 TRABALHOS FUTUROS

Existem outras ferramentas de ETL que poderiam realizar esta atividade de maneira similar ao que foi desenvolvido com o PDI, cabendo ao desenvolvedor a escolha do *software* de execução do estudo, portanto, estas ferramentas podem ser testadas em algum trabalho futuro. Uma opção alternativa é o software do Talend, que talvez possa trazer um tempo de carregamento das informações menor do que o constatado neste presente estudo.

Uma outra melhoria que foi identificada e pode ser trabalhada em um estudo futuro é a aplicação de um novo modelo de ETL em outros LMS, para

servir de fonte extração de informações com o intuito de prover uma maior integração com diferentes sistemas, providenciando dessa forma uma maior amplitude de usuários, devido ao aumento de sistemas de aprendizagem atendidos pelo *framework*.

REFERÊNCIAS

[1] PEREIRA, Alice Theresinha Cybis; SCHMITT, Valdenise; DIAS, M. R. A. C. Ambientes virtuais de aprendizagem**. AVA-Ambientes Virtuais de Aprendizagem em Diferentes Contextos. Rio de Janeiro: Editora Ciência Moderna Ltda**, p. 4-22, 2007.

[2] The world's most valuable resource is no longer oil, but data Disponível em: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. Acesso em: 10 mai. 2022

[3] MACIEL, A. **Framework FMDEV**. Disponível em: http://www.ecomp.poli.br/amam/software/fmdev Acesso em: 15 mai. 2022

[4] FERREIRA, João et al. **O processo etl em sistemas data warehouse**. In: INForum. 2010. p. 757-765.

[5] AGRAPETIDOU, Anna et al. An AutoML application to forecasting bank failures. **Applied Economics Letters**, v. 28, n. 1, p. 5-9, 2021.

[6] HUTTER, Frank; KOTTHOFF, Lars; VANSCHOREN, Joaquin. **Automated machine learning: methods, systems, challenges**. Springer Nature, 2019.

[7] MACÊDO, Pedro HR; SANTOS, Wylliams B.; MACIEL, Alexandre MA. **Análise de perfis de engajamento de estudantes de ensino a distância**. RENOTE, v. 18, n. 2, p. 326-335, 2020

[8] SILVEIRA, Malu Egídio da; JUSTI, Francis Ricardo dos Reis. Engajamento escolar: Adaptação e evidências de validade da escala EAE-E4D. **Psicologia: teoria e prática**, v. 20, n. 1, p. 110-125, 2018.

[9] XANTHOPOULOS, Iordanis et al. Putting the Human Back in the AutoML Loop. In: **EDBT/ICDT Workshops**. 2020.

[10] ANAND, Nitin; KUMAR, Manoj. Modeling and optimization of extraction-transformation-loading (ETL) processes in data warehouse: An overview. In: **2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)**. IEEE, 2013. p. 1-5.

[11] HARYONO, Edward Manopo et al. Comparison of the E-LT vs ETL method in data warehouse implementation: A qualitative study. In**: 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)**. IEEE, 2020. p. 115-120.

[12] ZODE, Madhu. The Evolution of ETL. **Retrieved on,** v. 6, n. 06.

[13] HATLE, DGNGS et al. Pentaho data integration tool. **Business Intelligence Tool**, p. 2-18, 2013.

[14] ZHOU, Wang; KLEIN, Levente J.; LU, Siyuan. Pairs autogeo: an automated machine learning framework for massive geospatial data. In: **2020 IEEE International Conference on Big Data (Big Data)**. IEEE, 2020. p. 1755-1763.

[15] ALNEGHEIMISH, Sarah et al. Cardea: An open automated machine learning framework for electronic health records. In: **2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)**. IEEE, 2020. p. 536-545.

[16] DA SILVA, Raniel Gomes. **Desenvolvimento de uma Solução de Aprendizado de Máquina Automatizado Integrável a Múltiplos Ambientes Virtuais de Aprendizagem.** Dissertação (Mestrado em Engenharia da Computação) – Universidade de Pernambuco, Recife, 2020

[17] ALVES, Lynn; BARROS, Daniela Melaré Vieira; OKADA, Alexandra. Moodle: estratégias pedagógicas e estudos de caso. 2009.

[17] Estrutura Redux escalável com Ducks. Disponível em: https://blog.rocketseat.com.br/estrutura-redux-escalavel-com-ducks/ Acesso em: 15 mai. 2022

[19] Introdução com bancos de dados no Subsistema do Windows para Linux. Disponível em: https://docs.microsoft.com/pt-br/windows/wsl/tutorials/wsl-database Acesso em: 15 mai. 2022