APLICAÇÃO DE TÉCNICAS DE ENGENHARIA DE DADOS PARA COMENTÁRIOS DE FEEDBACK E AVALIAÇÃO DE MELHORIA INSTITUCIONAL

Application of data engineering techniques for feedback comments and evaluation of institutional improvement

Aquiles Silva Aguiar
Graduando
Sistemas de Informação
aquilessilva@ucl.br

Gabriel de Souza Passos
Graduando
Sistemas de Informação
gabrielpassos@ucl.br

Marlon Ferrari
Orientador
Faculdade UCL
marlonferrari@ucl.br

A lista completa com informações de autoria está no final do artigo

RESUMO

Objetivo: Este artigo utiliza técnicas de engenharia de dados em *feedbacks* institucionais para criar um conjunto de informações para processamento de linguagem natural. Os dados utilizados são históricos de comentários de alunos de diferentes períodos e cursos, abordando avaliações específicas sobre áreas acadêmicas e infraestrutura. O objetivo é facilitar a interpretação de novas informações e obter insights relevantes a partir da base de dados construída.
Método: A partir do estudo de caso a uma base extraída a partir dos sistemas de avaliação, foram realizados levantamentos bibliográficos acerca das etapas sistêmicas que envolvem tratamentos de dados. Compreendendo as características das bases, foram realizados inventários de aplicações, técnicas e tecnologias propícias à realização de todas as etapas a fim de obter uma base funcional e orientada a processamento de linguagem natural, que possui suas características e requisitos.
Resultado: Criou-se um conjunto de dados contendo 2 colunas, de 2 bases diferentes unidas e relacionadas, tratadas pelos métodos do ETL (Extração, Transformação e Carregamento) e ao fim com 27.651 registros.
Conclusões: Com base nos testes realizados, foi confirmado que o *dataset* possui uma boa performance e pode ser aplicado com sucesso durante as tomadas de decisões na instituição. Isso ressalta a importância das etapas de transformação de dados nessa área, pois garantem a qualidade e confiabilidade dos dados utilizados. Ao realizar as transformações adequadas, é possível obter um dataset otimizado, resultando em análises mais precisas e embasadas para uma tomada de decisão eficaz.

PALAVRAS-CHAVE: Engenharia de dados, Meio Institucional, Feedback, Transformação, Tratamento, ETL, Informação.

ABSTRACT

Objective: This article uses data engineering techniques in institutional feedback to create an information set for natural language processing. The data used are historical comments from students from different periods and courses, addressing specific assessments on academic areas and infrastructure. The objective is to facilitate the interpretation of new information and obtain relevant insights from the constructed database.
Method: Based on the case study, a base extracted from the evaluation systems, bibliographic surveys were carried out on the systemic steps that involve data processing. Understanding the characteristics of the data, an inventory of applications, techniques and technologies conducive to carrying out all the steps was carried out in

order to obtain a functional base oriented to natural language processing, which has its characteristics and requirements.
Result: A dataset was created containing 2 columns, from 2 different bases united and related, treated by ETL methods (Extraction, Transformation and Loading) and at the end with 27,651 records.
Conclusions: Based on the tests carried out, it was confirmed that the dataset has a good performance and can be successfully applied. This underscores the importance of data transformation steps in this area, as they ensure the quality and reliability of the data used. By performing the appropriate transformations, it is possible to obtain an optimized dataset, resulting in more accurate and informed analyzes for effective decision making.

KEYWORDS: Data Engineering, Institutional Environment, Feedback, Transformation, Treatment, ETL, Information.

1 INTRODUÇÃO

A geração de dados vem crescendo exponencialmente com o passar dos anos. A coleta, armazenamento e processamento desses dados se tornaram desafios significativos para as organizações e companhias. Contudo, o processamento de dados brutos podem ser aprimorados através da aplicação de técnicas e tecnologias de engenharia de dados (REZENDE, 2003).

A pesquisa aborda a aplicação de técnicas e tecnologias de engenharia de dados presentes em fontes bibliográficas e de repositórios para dados aplicados ao processamento de linguagem natural, com ênfase nas etapas de coleta, retenção e processamento dos dados brutos, para geração de um conjunto de dados pré processados e apto ao desenvolvimento de modelos de linguagem natural.

Para a construção da ideia, foram levantados diversos pontos, e um deles foi o quão importante a etapa de ETL (Extração, Transformação e Carregamento), bem definida e estruturada é para uma organização. De acordo com Kimball e Ross esse processo:

> "... é responsável por converter, limpar e enriquecer os dados extraídos antes de serem carregados no destino final" (KIMBALL; ROSS; THORNTHWAITE; MUNDY; BECKER, 2013).

De acordo com Russell e Norvig (2013), a engenharia é uma disciplina que busca aplicar conhecimentos teóricos e práticos para resolver problemas complexos em diferentes áreas. Já a palavra “Dados” conceitua-se conforme Provost e Fawcett (2013) como informações coletadas e/ou geradas por sistemas, dispositivos ou usuários, que podem ser armazenados, processados e analisados para obter insights e conhecimentos.

Kelleher e Tierney dizem que:

> "O tratamento de dados envolve a manipulação e organização de dados para torná-los mais acessíveis e utilizáveis para análise" (KELLEHER; TIERNEY, 2018, p. 17).

A Engenharia de Dados em si é a área que atua no processamento e tratamento dos dados para a utilização dos mesmos, que futuramente pode ser usada para análises, como desenvolvimento de *dashboards*, até para uma simples exportação para uma planilha (SIRQUEIRA; DALPRA, 2018).

Como caso de uso do presente trabalho, busca-se junto à instituição de ensino Faculdade UCL, dados das avaliações institucionais de *feedback* mantidas pela CPA (Comissão Própria de Avaliação), onde encontram-se *feedbacks* de discentes acerca de eixos como infraestrutura, docentes e disciplinas.

Salles e Lopes (2021) mostram que a análise de sentimentos pode ser aplicada para avaliar a satisfação dos usuários em relação a produtos ou serviços. A classificação de tópicos pode ser utilizada para identificar os principais assuntos abordados nos comentários. A modelagem de tópicos pode ser aplicada para descobrir padrões e tendências nos comentários.

Tendo em vista a importância dos dados de *feedbacks* institucionais para a tomada de decisões e revisão de novos caminhos para a melhoria contínua, bem como os desafios que envolvem o pré processamento de dados para que estes tornem-se acionáveis, o presente trabalho busca integrar as técnicas de Engenharia de Dados para a disponibilização de um conjunto de dados aptos às tarefas de processamento de linguagem natural, com o propósito de facilitar o entendimento e mensuração dos comentários dos discentes, presentes nos dados brutos da avaliação institucional.

2 REFERENCIAL TEÓRICO

2.1 ETL

Seu conceito mais básico consiste na extração, tratamento e carregamento dos dados de alguma fonte, como apresentado na Figura 1. Esta etapa é presente no tipo de armazenamento chamado de *Data Warehouse*, no qual o primeiro passo para iniciar o processo de Engenharia de Dados é a própria coleta.

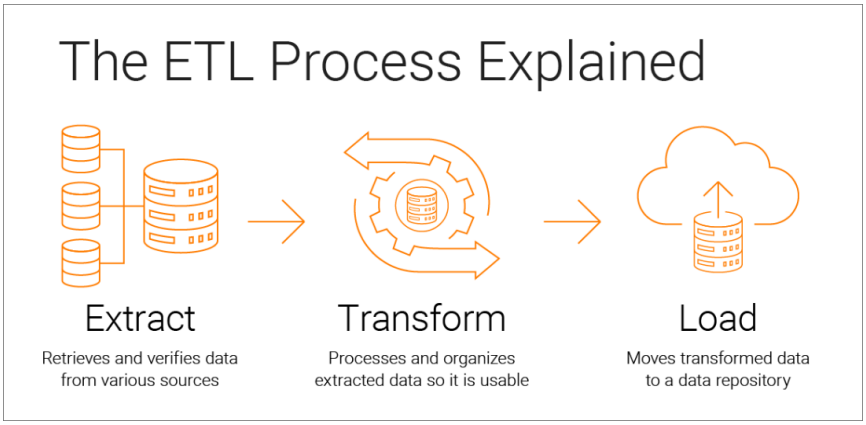

Figura 1. Processo de ETL - (Informatica, 2023)

No processo de extração, os dados podem ser consumidos de diferentes locais, como banco relacionais ou não relacionais, da *cloud* e arquivos de formatos distintos que deseja-se obter *insights*. Com isso, estes dados coletados precisam ser limpos, retirando inconsistências, como: campos nulos, caracteres especiais, dependendo do tipo de análise abordada no escopo do projeto, e outros casos que podem prejudicar a etapa de análise, parte conhecida como “transformação”.

Já na última etapa, a de carregamento, consiste na inserção dos dados que passaram pelo processo de transformação, em um tipo de armazenamento, como *Data Warehouses*, banco de dados e outros. E assim, após o tratamento e carregamento dos mesmos, podemos enfim utilizar os dados para gerar informações relevantes durante os processos de análise.

2.2 *DATA CLEANSING*

O *Data Cleansing* consiste na limpeza, transformação e padronização dos dados para torná-los mais acessíveis e utilizáveis para análise, garantindo a qualidade e confiabilidade.

De acordo com Kelleher e Tierney (2018), a manipulação e organização de dados são essenciais para torná-los mais acessíveis e utilizáveis para análise. Para alcançar esses objetivos, diversas técnicas e tecnologias podem ser usadas durante o processo de *data cleansing*, incluindo a detecção e correção de erros de digitação, padronização de dados, remoção de dados duplicados e/ou nulos.

Neste contexto, o presente artigo tem como objetivo principal a execução de um processo de data cleansing eficiente, visando a preparação de uma base de dados limpa e estruturada. Essa etapa é crucial para otimizar a visualização e manipulação dos dados, além de possibilitar a geração de pré-modelos que serão utilizados posteriormente em relatórios e análises mais detalhadas. Ao adotar uma abordagem cuidadosa de data cleansing, espera-se garantir a confiabilidade e qualidade dos dados, proporcionando uma base sólida para *insights* valiosos e tomadas de decisão embasadas.

2.3 *DATA LOADING*

O processo de *Data Loading* é fundamental para carregar dados de diversas fontes, prepará-los e transformá-los para que possam ser utilizados por sistemas de análise de dados. Segundo Provost e Fawcett (2013), dados são informações coletadas e/ou geradas por sistemas, dispositivos ou usuários, que podem ser armazenados, processados e analisados para obter conhecimento. Para que esses dados possam ser utilizados de maneira eficiente, é necessário que eles sejam estruturados, limpos e preparados para a análise. Após a preparação dos dados, eles são escritos em um formato de arquivo, para que possam ser utilizados. Dessa forma, é possível armazenar os dados preparados de maneira organizada e eficiente, facilitando o acesso e a análise posterior dos mesmos.

2.4 OLTP

De acordo com Connolly e Begg (2014) o *Online Transaction Processing* ou Processamento de Transações Online é utilizado para gerenciar transações em tempo real em sistemas de bancos de dados, com um foco em garantir a consistência e integridade dos dados (CONNOLLY; BEGG, 2014).

Para a execução de transações em bancos de dados operacionais em tempo real, o OLTP é o sistema chave. Um sistema de banco de dados OLTP tem como uma das características sua estrutura normalizada, que tem como sentido fazer a divisão de tabelas com o intuito de reduzir ocorrências de redundâncias (SILBERSCHATZ; KORTH; SUDARSHAN, 2010).

No contexto do artigo, a OLTP é a arquitetura onde se encontra a plataforma de avaliação institucional e as bases de dados transacionais, sendo fundamental sua a coleta dos dados.

2.5 OLAP

O Processamento Analítico Online, no inglês *Online Analytical Processing*, consiste na técnica para analisar de várias dimensões diante de um grande volume de dados em tempo real (HAN; KAMBER, 2006).

O seu processo aborda a construção de cubos de dados multidimensionais, que levam a capacidade de analisar os dados de várias perspectivas. Com isso, é possível perceber a presença de padrões, formatos e tendências. Por esse motivo essa técnica é muito utilizada por grandes empresas, sendo de grande importância também no suporte ao *Business Intelligence* (MICROSOFT, 2021).

No âmbito institucional, a análise multidimensional dos dados permite que seja identificado padrões que auxiliam na descoberta da problemática a qual requer atenção e que possa ser desenvolvido algum plano de ação para resolução do problema.

Os dados originados do OLAP são derivados do OLTP, e são geralmente guardados em *Data Warehouses* ou em *Data Lakes*, depois de serem tratados a partir de conceitos de Extração, Transformação e Carregamento dos dados, como é mostrado na Figura 2.

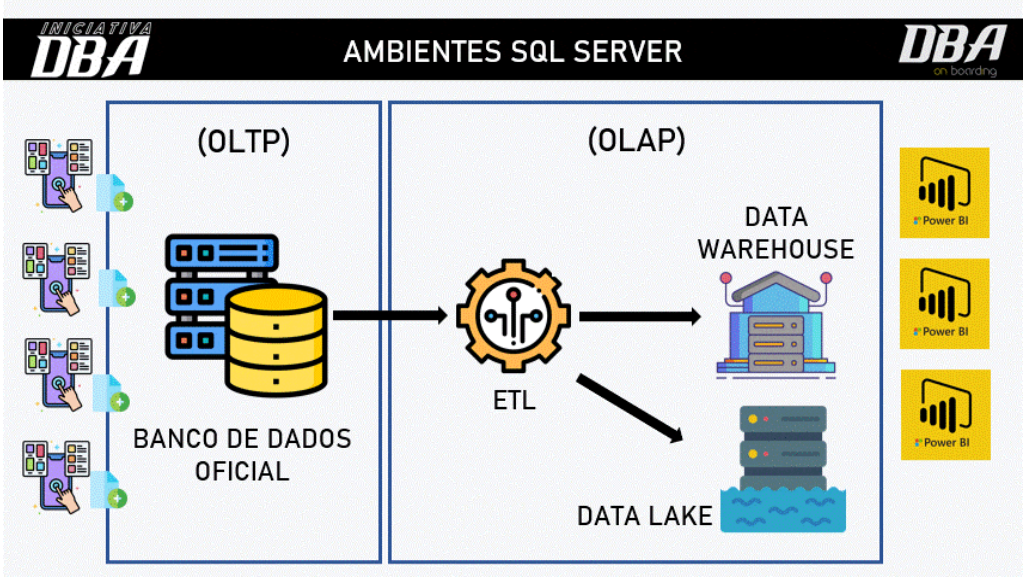

Figura 2. Processo OLTP e OLAP - (DBAonboarding, 2021)

No geral, a utilização destas técnicas foram de extrema importância no que se diz respeito a análise de comentários dos discentes, devido aos comentários estarem aptos a passar pelo processo de tratamento, depois armazenados e utilizados por ferramentas e mecanismos analíticos e de processamento de linguagem natural.

2.6 DESNORMALIZAÇÃO DE DADOS

Consiste no processo de mudança de estrutura de um repositório de dados normalizado com intuito à atuação das consultas e reduzir também a complexidade dos modelos analíticos. Esta técnica pode ser utilizada de diversas formas, como duplicação de informações, criação de identificadores e entre outros. Kimball e Ross (2013) mostram que a desnormalização é muito útil na hora de fazer consultas com um volume de dados médio e alto, e com um grande número de usuários. Porém também deve-se ter cuidado com o uso do mesmo, pois caso não seja utilizada de forma correta, pode resultar em inconsistências de dados e acarretar problemas na manutenção.

Esse processo pode ser inserido também para melhorar a performance de consultas com o fins analíticos (INMON, 2005). A partir disso, poderá ser criado estruturas de dados que farão análises mais complexas dos dados.

2.7 *PARQUET*

Desenvolvido pelo projeto Apache Hadoop, o *parquet* é um formato de arquivo de código aberto, no qual os dados são armazenados em colunas, isso significa que é muito útil quando é abordado questões como processamento de determinadas consultas e espaço de armazenamento (ALURA, 2021).

De acordo com White (2015) o *parquet* viria a ser cada vez mais utilizado com um formato para armazenamento de uma grande quantidade de dados, e é exatamente isso que vem acontecendo, pelo fato do conjunto de dados ser acessado frequentemente e pela execução de várias consultas que necessitam ser rápidas. Esse formato de arquivo é recorrentemente utilizado em áreas de *Big Data*, Engenharia de Dados e na Análise de Dados.

O *parquet* foi construído com o objetivo de ser compatível com vários sistemas de processamentos de dados, como *Hadoop* e o *Spark* tornando-se muito popular nas empresas que trabalham com um grande volume de dados.

Essas características tornam o *parquet* altamente vantajoso para o desenvolvimento do artigo em questão. Além disso, permite consultas mais ágeis em comparação com outros formatos, como o CSV (*Comma-Separated Values* ou Valores Separados por Vírgula), conforme ilustrado na Figura 3. Essa eficiência do *parquet* proporciona ganhos significativos em termos de desempenho e facilita a manipulação dos dados no contexto do estudo.

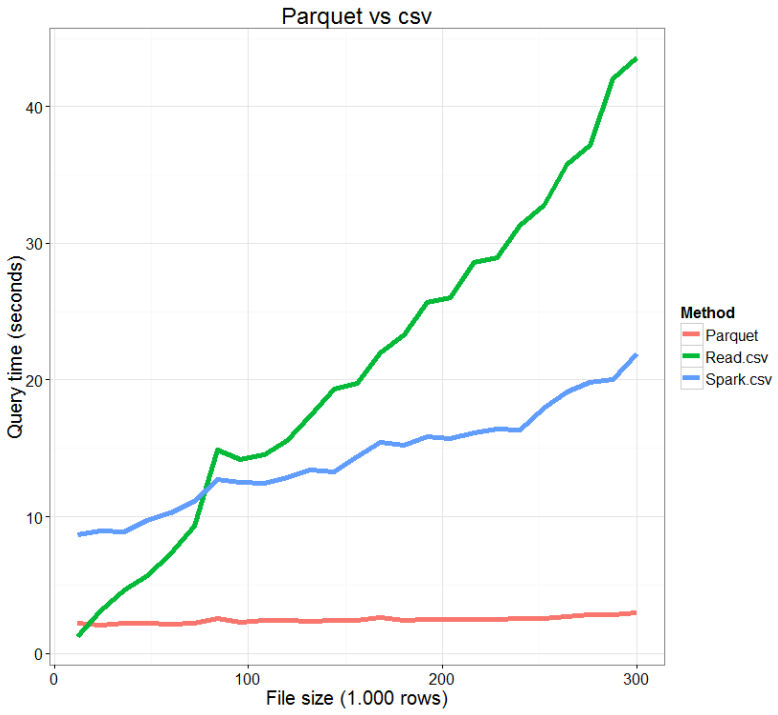

Figura 3.Comparativo Parquet x CSV x Spark CSV - (Dados Dados Dados, 2015)

Outro fator importante é a fácil integração deste formato com outras plataformas, sendo assim, de grande ajuda na Engenharia de Dados. Como resultado, tem-se uma melhoria nas etapas de carregamento, transformação e armazenamento dos comentários.

3 METODOLOGIA

A metodologia adotada neste trabalho baseou-se em pesquisas bibliográficas e artigos científicos para explicar e mostrar, com o desenvolvimento de um projeto, como as etapas de engenharia de dados podem impactar de forma positiva para a formação de um pré modelo de dados para geração de resultados, por meio de ferramentas de análise de sentimentos no ambiente institucional.

3.1 FERRAMENTAS UTILIZADAS

Para confecção do atual trabalho, foi necessário a utilização de algumas ferramentas que foram de suma importância nos processos de coleta, tratamento e armazenamento. Elas foram escolhidas a partir de suas eficácia e eficiência no escopo de como seriam utilizadas no projeto. As ferramentas apresentadas no Quadro 1, foram fundamentais para o sucesso da pesquisa.

Quadro 1 – Painel Ferramentas Utilizadas

Ferramenta	Descrição
PySpark	É um *framework* para processamento de dados em larga escala, fundado pela empresa Apache Software Foundation Foi a chave para se trabalhar com o grande volume de comentários (DRABAS; LEE, 2017).
Python	Python é uma linguagem de programação de alto nível, interpretada e de propósito geral. É conhecida por sua sintaxe simples e legibilidade, o que a torna uma ótima opção para iniciantes em programação(YUILL; HALPIN, 2006).
Databricks	Plataforma que facilita a implementação e o treinamento de modelos de máquinas e análise de dados e também oferece a disponibilidade de utilizar em seu ambiente de forma dinâmica com diferentes linguagens (ALURA, 2023).
Pandas	O pandas é uma biblioteca em Python usada para análise e manipulação de dados. Fornece estruturas de dados eficientes, para trabalhar com conjuntos de dados tabulares. É possível carregar, transformar, filtrar e visualizar dados (MCKINNEY, 2011).

Fonte: Do Autor (2023)

3.2 DADOS COLETADOS

Foram exportadas a partir da base transacional do sistemas e avaliações institucionais seis tabelas, incluindo bases de modelo da avaliação, opções, perguntas, respostas e períodos. Dados

pessoais foram removidos durante esta etapa, para evitar identificação de professores e alunos envolvidos.

A Figura 4 ilustra as interações entre as tabelas "cpa_perguntas" e "cpa_resposta". Essas tabelas foram exportadas da base transacional do sistema e das avaliações institucionais, com dados pessoais removidos para preservar a privacidade dos envolvidos. Elas foram selecionadas especificamente para a análise dos dados das avaliações da faculdade, sendo cruciais para os tratamentos e modelos de análise subsequentes.

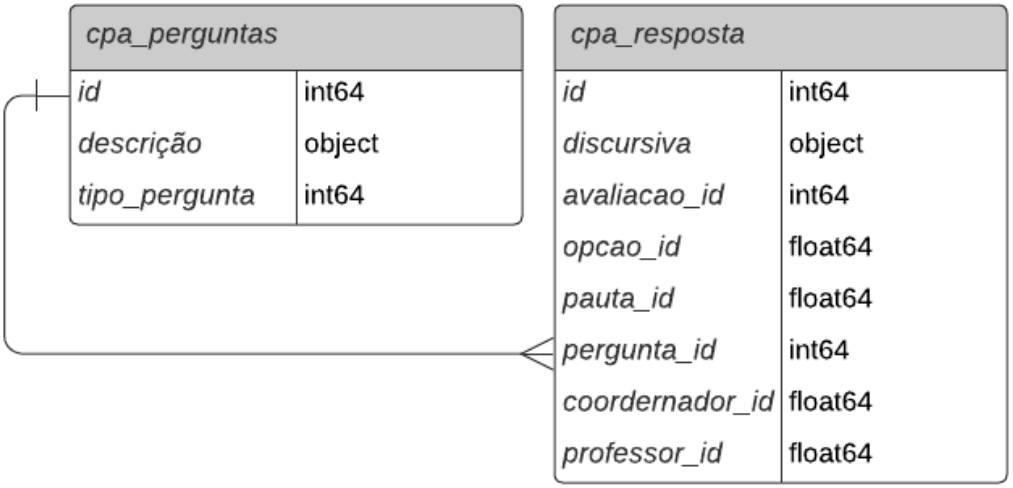

Figura 4.Modelo MER completo dos csv brutos - (Do autor, 2023)

Após o carregamento das tabelas na plataforma *Databricks*, foi iniciado o processo de verificação dos dados brutos, para as etapas iniciais de consistência e transformação de dados. A partir disso, montou-se um pequeno resumo com algumas especificações, para uma melhor visualização, com a quantidade de dados das duas tabelas somados antes do tratamento, como mostrado na Tabela 1.

Tabela 1 – Painel de Resumo dos Dados Brutos

Total de registros da base original	Total de registros únicos (por Ids)	Total de comentários não nulos	Total de comentários em brancos
546.013	546.013	27.651	518.196

Fonte: Do Autor (2023)

Depois de feita a extração e validação dos dados brutos, iniciou-se o processo de escolha da codificação da linguagem, este consiste em consolidar qual melhor tipo de caractere nas tabelas para se trabalhar. Tendo em mente este conceito, foi adotado a codificação ‘latin-1’, pelo fato da mesma identificar caracteres especiais e com acento.

3.3 ABORDAGEM EM ENGENHARIA DE DADOS

A Engenharia de Dados é uma área de conhecimento que atua na implementação de técnicas de processamento para gerenciar grandes escalas de dados, ela desempenha um papel crucial na qualidade dos dados e na preparação dos mesmos para análises e possíveis tomadas de decisões (MARZ; WARREN, 2015).

Como visto anteriormente no tópico 2.4 do referencial teórico, a Engenharia de Dados possui alguns processos, sendo eles o carregamento dos dados, o tratamento dos mesmos e por fim o seu carregamento. A partir destes conceitos, foi realizado o desenvolvimento do código com a finalidade de gerar um um modelo consistente para ser utilizado como fonte de análises.

3.3.1 *EXTRAÇÃO*

Inicialmente, com os dados extraídos das tabelas, foram gerados dois arquivos *CSV,* a partir das tabelas de perguntas e respostas da avaliação. Estes arquivos foram disponibilizados no *google drive.* Para a utilização desses dados, foi necessário carregá-los na plataforma *databricks*. Houve a utilização da biblioteca do python gdown, possibilitando realizar o *downloads* dos arquivos via *google drive* apenas com os id 's dos arquivos, como mostra a Figura 5.

```
#CPA_PERGUNTAS
!gdown --id 1-D4Uj-Ua1dlVxKJd58gI87UjlG-0M9aD

#CPA_RESPOSTAS
!gdown --id 1_ecvCrBpMXpSxdJnO4HHQnCDYHkIch3u
```

Figura 5. Download dos arquivos do google drive - (Do autor, 2023)

Em seguida, realizou-se a leitura e transformação dos dados para um formato de arquivo mais performático que o CSV, o *parquet*. Para desempenhar este papel, optou-se pela utilização da biblioteca pandas do python, que desempenha um bom papel para a manipulação de dados. Sendo assim, é possível efetuar a leitura dos arquivos, usando a codificação 'latin-1', para fazer com que os conteúdos dos mesmos não fossem corrompidos. A leitura dos arquivos e sua codificação pode ser vista na Figura 6.

```
import pandas as pd

cpa_respostas = pd.read_csv("/databricks/driver/eies_cpa_respostas_anon.csv",
    sep=";", encoding="latin-1")

cpa_perguntas = pd.read_csv("/databricks/driver/eies_cpa_perguntas.csv",
    sep="|")
```

Figura 6. Leitura dos arquivos csv - (Do autor, 2023)

Vale ressaltar que existem várias codificações, cada uma com sua especificidade para cada caso, sendo a mais utilizada a ‘utf-8’, mas sua aplicabilidade aos dados consumidos neste estudo não foi adequada conforme o esperado, a leitura dos arquivos pode ser vista na Figura 6.

Após os dados terem sido lidos, foram armazenados em um sistema de arquivos distribuído montado em um ambiente de trabalho *Databricks*, tendo os arquivos salvos em formato *parquet*. Após essa escrita, é possível utilizar os dados de forma global no sistema, já no formato escolhido, de forma bem mais performática, como mostra a Figura 7.

```
cpa_respostas.write.mode("overwrite").format("delta").parquet(
  "dbfs:///tcc/cpa_respostas.parquet")

cpa_perguntas.write.mode("overwrite").format("delta").parquet(
  "dbfs:///tcc/cpa_perguntas.parquet")
```

Figura 7. Salvamento em *parquet* - (Do autor, 2023)

Por fim, os dados unificados são armazenados em um sistema que leva em consideração aspectos de segurança e escalabilidade, permitindo o armazenamento e recuperação eficientes de dados para processamento e análise no futuro. A consolidação de dados é de suma importância na preparação dos dados de revisão institucional, que envolve extração, formatação e armazenamento de dados. Esta etapa é fundamental para garantir a qualidade, integridade e acessibilidade dos dados em preparação para tratamentos posteriores.

3.3.2 *DATA CLEANSING*

A partir da leitura dos arquivos em que os dados foram salvos no final da consolidação, como retratado na Figura 8, iniciou-se o processo de levantamento dos valores nulos e a verificação, em cada uma das duas tabelas, de quais colunas não seriam úteis para a futura análise.

```
cpa_perguntas = spark.read.options(header='true',
inferSchema='true').parquet("dbfs:///tcc/cpa_perguntas.parquet")
cpa_respostas = spark.read.options(header='true',
inferSchema='true').parquet("dbfs:///tcc/cpa_respostas.parquet")
```

Figura 8. Leitura dos arquivos consolidados - (Do autor, 2023)

Na validação das colunas, notou-se que na tabela “cpa_perguntas” seria necessário o uso de todas as colunas, pois era essencial saber qual o identificador da pergunta, qual a descrição e qual o identificação da resposta. Contudo, na outra base de dados chamada “cpa_respostas”, houve a exclusão das colunas "pauta_id", "coordenador_id", "professor_id", "avaliacao_id" e "opcao_id", como ilustrado na Figura 9, pois não seriam necessárias.

```
cpa_respostas = cpa_respostas.drop("pauta_id", "coordenador_id", "professor_id",
"avaliacao_id", "opcao_id")
```

Figura 9. Exclusão de algumas colunas da tabela “cpa_respostas” - (Do autor, 2023)

Após a verificação dos valores nulos na tabela “cpa_respostas” constatou-se que 94,1% das respostas eram questões discursivas vazias, e apenas 5,1% foram preenchidas pelos alunos, como apresentado na Figura 10.

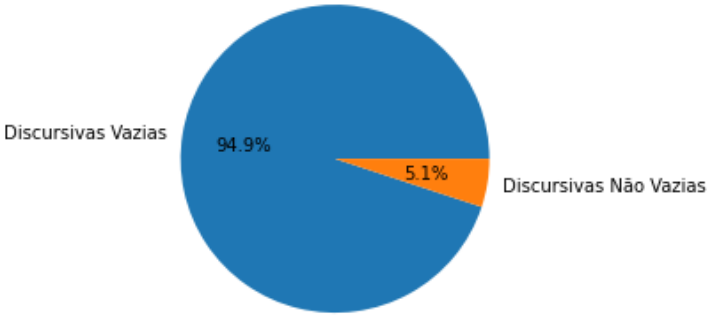

Figura 10.Gráfico das questões discursivas tabela “cpa_respostas” - (Do autor, 2023)

Foram removidos os caracteres especiais que estavam na coluna "discursiva" da tabela "cpa_respostas". O objetivo da remoção é apenas manter os caracteres alfanuméricos para garantir a qualidade e a consistência dos dados. A limpeza dos dados contribuiu para deixar a tabela mais compacta e pronta para análises subsequentes, sem interferências desnecessárias, como mostra a Figura 11.

```
cpa_respostas = cpa_respostas.withColumn("discursiva", regexp_replace("discursiva",
"[^\w\sáàâãéèêíïóôõöúüçñÁÀÂÃÉÈÊÍÏÓÔÕÖÚÜÇ]", ""))
```

Figura 11. Remoção de caracteres especiais presentes na coluna "discursiva" da tabela "cpa_respostas" - (Do autor, 2023)

A seguir, realizou-se a retirada de todos os dados nulos, valores vazios e caracteres especiais que estavam presentes na tabela como mostrado no código da Figura 6.3. Contudo, a exclusão dos nulos é feita após a remoção dos caracteres especiais, já que pode haver campos contendo apenas um caractere especial e depois de tratado, pode tornar-se vazio. Assim, após este tratamento a tabela irá ficar mais compacta e com as peças chaves para se continuar o projeto, como mostra a Figura 12.

```
def remove_empty_spaces(value):
    if value is None or value.strip() == "":
        return None
    else:
        return value.strip()

remove_empty_spaces_udf = udf(remove_empty_spaces, StringType())

cpa_respostas = cpa_respostas.withColumn("discursiva",
remove_empty_spaces_udf("discursiva"))

cpa_respostas = cpa_respostas.filter((cpa_respostas.discursiva != "null") |
(cpa_respostas.discursiva != "") | cpa_respostas.discursiva.isNotNull())
```

Figura 12. Retirada dos valores nulos e vazios na tabela “cpa_respostas” - (Do autor, 2023)

Com isso, o objetivo que a etapa de *data cleansing* tem é de garantir a qualidade e a consistência dos dados. Obtém-se por meio da identificação e remoção de valores nulos, exclusão de colunas irrelevantes e padronização dos dados.

3.3.3 *TRANSFORMAÇÃO DE DADOS*

Com o intuito de colocar os dados no melhor formato possível para serem trabalhados de forma eficiente, segue-se para a etapa de transformação de dados. Nela é feito o levantamento das tipagens dos mesmos por meio de consultas a um painel de dados para garantir a integridade e a consistência dos mesmos (HAN; KAMBER; PEI , 2011).

Como exposto na Figura 13 e na Figura 14, constatou-se que, antes das remoções das colunas, os dados foram obtidos da seguinte forma na tabela resposta e pergunta, respectivamente:

```
Out[139]: id                int64
discursiva         object
avaliacao_id        int64
opcao_id          float64
pauta_id          float64
pergunta_id         int64
coordenador_id    float64
professor_id      float64
dtype: object
```

Figura 13.Tipagem dos Dados na base “cpa_respostas” - (Do autor, 2023)

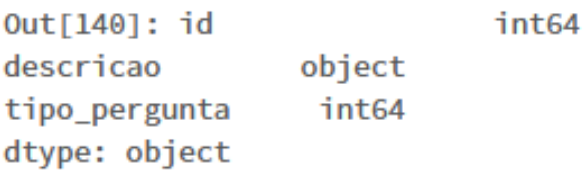

Figura 14.Tipagem dos Dados na base “cpa_perguntas” - (Do autor, 2023)

Após a análise inicial dos dados, identificou-se a necessidade de realizar transformações em determinados formatos de dados. Seguindo as práticas recomendadas, foram realizadas as tipagens dos dados, visando garantir sua integridade e consistência. Essas tipagens foram aplicadas aos dados resultantes da etapa de limpeza, conforme ilustrado na Figura 15 e Figura 16. Essas transformações permitiram adequar os dados a formatos apropriados para análises posteriores e possibilitaram a correta interpretação e manipulação dos mesmos.

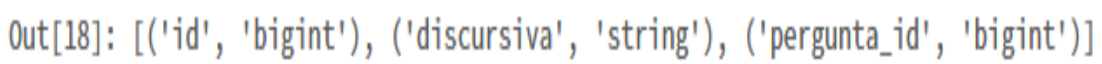

Figura 15 .Tipagem dos Dados na base “cpa_respostas” tratados e com remoção das colunas - (Do autor, 2023)

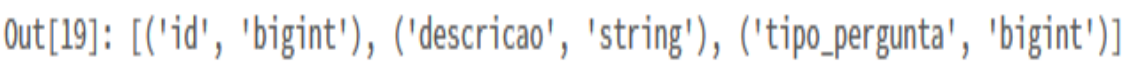

Figura 16.Tipagem dos Dados na base “cpa_perguntas” tratadas e com remoção das coluna - (Do autor, 2023)

A partir disso, obteve-se a possibilidade de avançar para a próxima etapa, a junção das tabelas com seus formatos de dados ajustados e a desnormalização.

3.3.4 *UNIFICAÇÃO DE DADOS*

Após a tipagem dos dados ser tratada, as tabelas foram combinadas para obter um *dataset* desnormalizado contendo todas as informações consolidadas em uma única tabela, conforme ilustrado na Figura 17. O objetivo dessa operação é consolidar os dados, proporcionando uma visão abrangente e integrada que facilite análises e extração de *insights* (HAN; KAMBER; PEI , 2011).

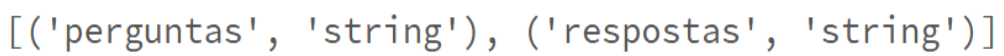

Figura 17. Representação da tipagem de dado da tabela desnormalizada - (Do autor, 2023)

Ao unir as tabelas, é possível estabelecer relações entre os dados com base em chaves de ligação, como identificadores únicos. Isso possibilita explorar conexões e padrões que podem não ser aparentes quando os dados estão separados em tabelas distintas.

4 RESULTADOS

A partir da confecção do trabalho, foram consolidados resultados relevantes com a aplicação das técnicas de Extração, Transformação e Carregamento nos conjuntos de dados oriundos das avaliações de *feedback* institucional. É importante ressaltar que os resultados apresentados são fruto de um processo contínuo de tratamento e análise de dados, e que novas avaliações e *feedbacks* devem ser constantemente incorporados para acompanhar as mudanças e garantir a qualidade do conjunto de dados.

4.1 MELHORIA NA QUALIDADE DOS DADOS

Através dos processos de ETL, foi possível realizar a padronização, a limpeza e organizar os dados da melhor forma. Isso resultou em bases sólidas para realizar as análises e tomada de decisões, pois será gerado dados mais confiáveis e mais consistentes.

A partir da Figura 18, é possível observar como o volume de dados foi severamente reduzido, mostrando assim, que grande parte dos mesmos eram nulos ou não eram de relevante importância para a presente pesquisa. Esses dados levantados foram as respostas discursivas, por serem valores mais ricos de informações para o escopo atual.

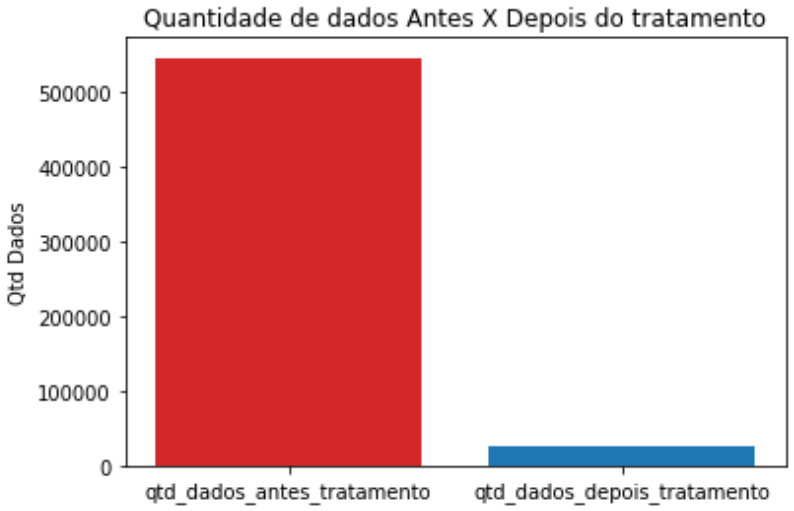

Figura 18 .Gráfico de Quantidade de dados Antes X Depois do tratamento - (Do autor, 2023)

4.2 FACILIDADE DE ACESSO E ANÁLISE DOS DADOS

Foi possível fornecer os dados preparados em um formato adequado para análise. Os dados foram transformados em um modelo estruturado, que pode ser acessado por parte dos gestores e tomadores de decisão pela utilização de ferramentas analíticas.

Em relação ao tipo de arquivo para armazenamento, apresenta-se na Figura 19 uma comparação entre os tempos de execução dos datasets em CSV e em *parquet*. A partir da análise, é possível verificar que o CSV possui um desempenho inferior, sendo portanto, o formato *parquet* muito mais eficiente no armazenamento de dados analíticos.

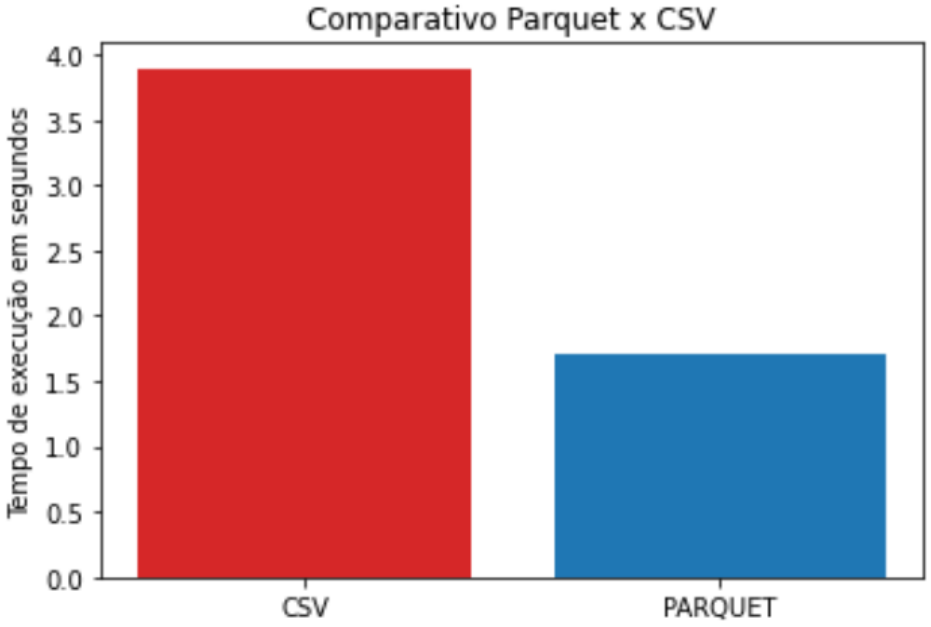

Figura 19.Gráfico de desempenho *Parquet* x CSV - (Do autor, 2023)

4.3 MODELOS ANALÍTICOS

Utilizando o pré modelo de dados gerado, foi possível desenvolver dois exemplos de vertentes que podem ser direcionados para a análise de dados. O primeiro exemplo baseia-se na identificação de domínio, mostrado na Figura 20.


```
{
    "sequence": "Muita coisa pra ela fazer Podia dividir as coisas pra
      ela nao ter que fazer tudo sozinha",
    "labels": [
        "feedback",
        "desempenho",
        "professor",
        "aulas",
        "infraestrutura"
    ],
    "scores": [
        0.4305402934551239,
        0.36636507511138916,
        0.08938682824373245,
        0.06821083277463913,
        0.04549692943692207
    ]
},
```


Figura 20. Estrutura da Identificação de Domínio - (Do autor, 2023)

Ela compreende na verificação de todos os 27.179 comentários, que foram obtidos após o tratamento, e na pontuação dos mesmos, na seção "*score",* com base nas *labels* inseridas. A *label* é uma palavra, rótulo ou tema que é definido antes do início da análise, a qual é o parâmetro principal para a verificação e a identificação dos comentários.

Já na Figura 21, foi gerado um gráfico com o intuito de se ter um panorama geral de como o contexto dos comentários estão divididos nas tags anteriores.

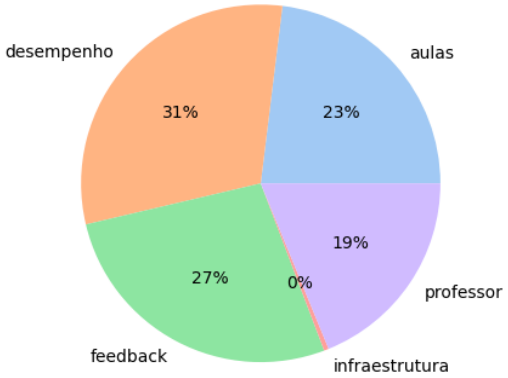

Figura 21. Gráfico geral da Identificação de Domínio - (Do autor, 2023)

Depois seguiu-se para o outro exemplo de insight com o pré modelo, a análise de sentimento. Ela tem como fundamento verificar como uma frase pode ser avaliada a partir da polaridade de cada palavra. A polaridade baseia-se na classificação de uma palavra entre os pólos bom, ruim ou neutro, tendo uma variação de 0 a 1, como mostrado na Figura 22.


```
{
    "sequence": "Muita coisa pra ela fazer Podia dividir as coisas pra
      ela nao ter que fazer tudo sozinha",
    "labels": [
        "negativo",
        "neutro",
        "positivo"
    ],
    "scores": [
        0.519202709197998,
        0.3244490623474121,
        0.156348317861557
    ]
},
```


Figura 22. Estrutura da Análise de Sentimento- (Do autor, 2023)

E por fim na Figura 23, tem-se a visão a partir de uma amostragem de 5.000 comentários, de uma total de 27.179 da base dados, tendo eles sua classificação de acordo com suas polaridades.

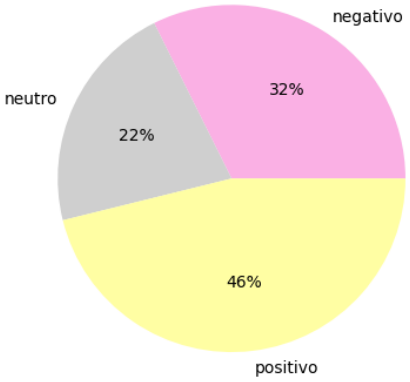

Figura 23. Gráfico geral da Análise de Sentimentos- (Do autor, 2023)

4.4 SUPORTE À TOMADA DE DECISÕES ESTRATÉGICAS

O modelo desnormalizado compreende uma fonte de dados importante às decisões, de forma que este contribui diretamente para a construção dos modelos analíticos. Para o caso de ser adotado em tarefas de processamento de linguagem natural, este já se encontra com boa performance e com as colunas aptas a servirem de base ao motor de treinamento dos modelos.

Adicionalmente, todas as etapas de transformação encontram-se em linguagem *Spark* e compatível com processamentos agendados e periódicos, facilitando a integração por quaisquer plataformas analíticas. Este é um importante resultado, pois cria a possibilidade de integração com a plataforma transacional da avaliação institucional, que será discutida na seção posterior.

5 TRABALHOS FUTUROS

De acordo com com o tema, notou-se a possibilidade de exploração de algumas áreas de conhecimento para trabalhos futuros. A principal delas é a implementação de processamento de linguagens naturais, incluindo a análise de sentimentos. A partir disso, seria possível realizar uma análise mais aprofundada dos comentários, podendo assim, incluir um gráfico de satisfação de acordo com as avaliações dos usuários, como mostrado no exemplo da Figura 24.

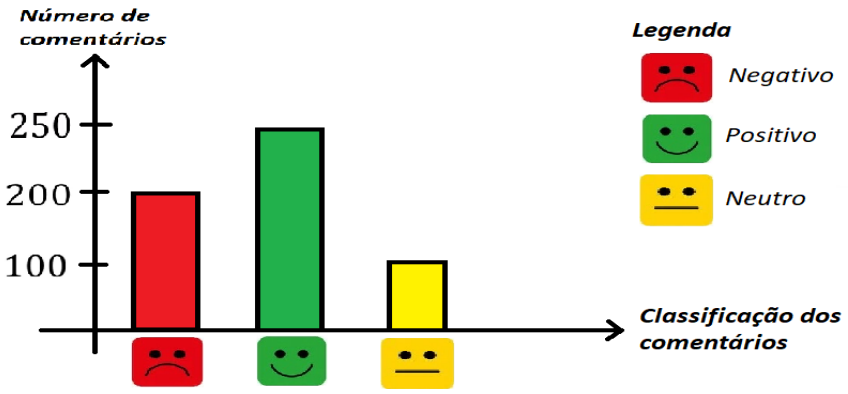

Figura 24. Exemplo do modelo gráfico depois da análise de sentimento - (Do autor, 2023)

Outra melhoria a ser implementada será a adição de um escopo maior antes da criação do pré-modelo de dados, a qual será adicionado os campos com as avaliações objetivas e também a divisão por períodos de cada curso. Isso fará com que seja possível um levantamento mais detalhado, amplo e assertivo no momento de coletar *insights*.

Em relação à Engenharia de Dados, seria possível agregar também dados de redes sociais, ambiente online de aprendizagem para maior amplitude da avaliação institucional, possibilitando a construção de modelos mais assertivos para compreensão de contextos e pleitos dos discentes.

Por fim, ter possibilidade de explorar a integração dos processos de ETL desenvolvidos neste estudo com a plataforma principal da faculdade. Essa integração permitiria a atualização contínua dos dados após as avaliações de *feedback*, possibilitando um acompanhamento em tempo real da qualidade do ensino e identificando áreas de melhoria. Ao manter os dados sempre atualizados e disponíveis na plataforma, os responsáveis pela gestão acadêmica teriam acesso imediato às informações relevantes, facilitando a tomada de decisões estratégicas e a implementação de ações de melhoria de forma ágil e eficaz. Dessa forma, o processo de ETL se tornaria uma peça fundamental no ciclo contínuo de análise e aprimoramento da qualidade educacional na instituição.

6 CONSIDERAÇÕES FINAIS

Neste estudo, foram aplicadas técnicas de ETL (Extração, Transformação e Carga) para coletar, limpar e integrar os dados de feedback dos estudantes. Isso proporcionou uma base sólida para análises posteriores, que irá resultar em insights valiosos para identificar áreas de melhoria na instituição de ensino.

A utilização de uma metodologia estruturada e apropriada para o tratamento e integração dos dados permitiu obter uma visão mais abrangente das informações, embasando a tomada de decisões estratégicas. Essas técnicas revelaram oportunidades de melhoria em processos internos e práticas pedagógicas, contribuindo para promover uma educação de qualidade.

Em resumo, este estudo evidencia o potencial das técnicas de ETL na análise de dados de feedback dos estudantes. A abordagem estruturada adotada permitiu identificar insights relevantes, que orientaram melhorias na qualidade do ensino e na experiência dos estudantes. A aplicação adequada dessas técnicas foi essencial para o sucesso do estudo e para o contínuo aprimoramento da instituição como um todo.

REFERÊNCIAS

ALURA. **Arquivos Parquet**. Disponível em: https://www.alura.com.br/artigos/arquivos-parquet. Acesso em: 03 abr. 2023.

ALURA. **Databricks: o que é e para que serve?** Paulo Calanca,2023, Disponível em: https://www.alura.com.br/artigos/databricks-o-que-e-para-que-serve. Acesso em: 12 jun. 2023.

BATISTA, G. e Monard, M. (2003). **Preprocessing data for data mining**. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

CARLSSON, Kjell; D, Ph; Gualtieri, Mike.The Forrester Wave™: **Notebook-Based Predictive Analytics And Machine Learning Solutions, Q3 2018**. Forrester Research, 2018. E-book. Disponível em: https://www.oracle.com/a/ocom/docs/dc/res143219-lpd100744077.pdf?elqTrackId=08c8e32cb9df4a52b516673b7ab3f638&elqaid=75081&elqat=2. Acesso em: 04 mai. 2023.

CONNOLLY, T. e Begg, C. (2014). **Bancos de Dados: Projeto, Implementação e Gerenciamento (4ª edição). LTC,** Pearson.

DADOS DADOS DADOS. **Comparativo Parquet x CSV x Spark**. Disponível em: https://dadosdadosdados.wordpress.com/2015/12/30/benchmarking-csv-vs-parquet/. Acesso em: 03 mai. 2023.

DBAONBOARDING. **Processo OLTP e OLAP**. Disponível em: https://static.wixstatic.com/media/20d4f8_842daf80cd994c3e822ff370e3df0e96~mv2.gif. Acesso em: 30 abr. 2023.

DRABAS, Tomasz; LEE, Denny. **Learning PySpark.** Birmingham, UK: Packt Publishing Ltd, 2017.

HAN, J., Pei, J. e Kamber, M. (2011). **Data mining: concepts and techniques (3rd ed.)**. San Francisco, CA: Morgan Kaufmann Publishers.

HAN, J. e Kamber, M. (2006). **Data mining: concepts and techniques (2nd ed.)**. San Francisco, CA: Morgan Kaufmann Publishers.

INFORMÁTICA. **Processo de ETL**. Disponível em: https://www.informatica.com/tw/resources/articles/what-is-etl.html. Acesso em: 30 abr. 2023.

INMON, W. H. (2005). **Building the Data Warehouse**. Indianapolis, Indiana: John Wiley & Sons.

KELLEHER, John; Brendan, Tierney. **Data Science.** Cambridge MA: MIT Press, 2018. E-book. Disponível em: https://encurtador.com.br/prwyP. Acesso em: 03 mai. 2023.

KIMBALL, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2013). **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.** John Wiley & Sons.

MARZ, N., & WARREN, J. (2015). **Big Data: Principles and best practices of scalable real-time data systems**. New York, Manning Publications.

MCKINNEY, Wes et al. **pandas: a foundational Python library for data analysis and statistics**. Python for high performance and scientific computing, v. 14, n. 9, p. 1-9, 2011. Disponível: https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf. Acesso em: 12 jun. 2023.

MICROSOFT. **Azure Databricks - External Data - Parquet**. Disponível em: https://learn.microsoft.com/pt-br/azure/databricks/external-data/parquet. Acesso em: 16 de abril de 2023.

MICROSOFT. **Visão geral do OLAP (Online Analytical Processing)**. Disponivel em: https://support.microsoft.com/pt-br/office/vis%C3%A3o-geral-do-olap-online-analytical-processing-15d2cdde-f70b-4277-b009-ed732b75fdd6. Acesso em : 30 abr. 2023.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business: What you need to know about data mining and data-analytic thinking**. " O'Reilly Media, Inc.", 2013. Disponível em: https://encurtador.com.br/pvCK2. Acesso em: 24 abr. 2023.

PYTHON, Why. **Python**. Python Releases Wind, v. 24, 2021. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1f2ee3831eebfc97bfafd514ca2abb7e2c5c86bb. Acesso em: 12 jun. 2023.

REZENDE, Solange Oliveira et al. **Mineração de dados. Sistemas inteligentes: fundamentos e aplicações**, v. 1, p. 307-335, 2003. Acesso em: 24/05/2023.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach (3rd ed.)**. “Pearson”, 2003. Disponível em: http://git-alunos.lis.ulusiada.pt/courses/ia2021/handouts/labs/lab1.pdf. Acesso em: 24 abr. 2023.

SILBERSCHATZ, A., Korth, H. F. e Sudarshan, S. (2010). **Sistema de Bancos de Dados (6ª edição)**. McGraw Hill.

SIRQUEIRA, Tassio; DALPRA, Humberto. **NoSQL e a Importância da Engenharia de Software e da Engenharia de Dados para o Big Data.** Sociedade Brasileira de Computação, 2018. Acesso em 10 abr. 2023.

SPARK. **Spark Overview**. Estados Unidos: Apache Spark, 2023. Disponível em: https://spark.apache.org/docs/latest/#spark-overview. Acesso em: 04 mai.2023.

WHITE, T. (2015). **Hadoop: The Definitive Guide**. Sebastopol, CA: O'Reilly Media, Inc.

NOTAS

AGRADECIMENTOS
Não se aplica.

CONTRIBUIÇÃO E AUTORIA
Co-orientação: Marlon Ferrari (Professor da Faculdade UCL Manguinhos).

FINANCIAMENTO
Não se aplica.

CONSENTIMENTO DE USO DE IMAGEM
Não se aplica.

CURSO
Sistemas de Informação

COORDENADOR DO CURSO
André Ribeiro da Silva

DATA DE ENTREGA
16/06/2023

BANCA AVALIADORA
André Ribeiro da Silva
Igor Martins Dessaune

DECLARAÇÃO DE INEXISTÊNCIA DE PLÁGIO
Declaro que o trabalho não contém plágio ou autoplágio total ou parcial. Todo o conteúdo utilizado como citação direta ou indireta foi indicado e referenciado.