UNIVERSIDADE FEDERAL DO CEARÁ

INSTITUTO UNIVERSIDADE VIRTUAL

CURSO DE GRADUAÇÃO EM SISTEMAS E MÍDIAS DIGITAIS

FRANCISCO EDVAN DE OLIVEIRA JUNIOR

DA COLETA À ANÁLISE: UM ESTUDO DE CASO USANDO DADOS PÚBLICOS DE SAÚDE, SEGURANÇA E DESENVOLVIMENTO SOCIAL.

FORTALEZA

2022

FRANCISCO EDVAN DE OLIVEIRA JUNIOR

DA COLETA À ANÁLISE: UM ESTUDO DE CASO USANDO DADOS PÚBLICOS DE SAÚDE, SEGURANÇA E DESENVOLVIMENTO SOCIAL.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em SISTEMAS E MÍDIAS DIGITAIS do INSTITUTO UNIVERSIDADE VIRTUAL da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em SISTEMAS E MÍDIAS DIGITAIS.

Orientadora: Profª. Dra. Ticiana Linhares Coelho da Silva

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O47c Oliveira Junior, Francisco Edvan de.
Da coleta à análise : um estudo de caso usando dados públicos de saúde, segurança e desenvolvimento social / Francisco Edvan de Oliveira Junior. – 2022.
40 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Instituto UFC Virtual, Curso de Sistemas e Mídias Digitais, Fortaleza, 2022.
Orientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. Dados Públicos. 2. KDD. 3. Análise de Dados. 4. Banco de dados. I. Título.

CDD 302.23

FRANCISCO EDVAN DE OLIVEIRA JUNIOR

DA COLETA À ANÁLISE: UM ESTUDO DE CASO USANDO DADOS PÚBLICOS DE SAÚDE, SEGURANÇA E DESENVOLVIMENTO SOCIAL.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em SISTEMAS E MÍDIAS DIGITAIS do INSTITUTO UNIVERSIDADE VIRTUAL da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em SISTEMAS E MÍDIAS DIGITAIS.

Aprovada em: 14/07/2022

BANCA EXAMINADORA

Profª. Dra. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Leonardo Oliveira Moreira
Universidade Federal do Ceará (UFC)

M.e Lucas Peres Gaspar
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

À minha mãe, Maria Juliana, por todo o suporte prestado, apoio, amizade e amor nos momentos em que precisei. Sem sua dedicação ao longo da vida, nenhuma conquista seria possível.

À instituição Universidade Federal do Ceará, por todo conhecimento a mim possibilitado durante meu tempo de graduação. É um orgulho imensurável fazer parte do corpo discente desta instituição.

À minha orientadora Dra. Ticiana Linhares Coelho da Silva, por toda a orientação prestada no desenvolvimento deste trabalho.

“O recurso mais valioso do mundo não é mais o petróleo, mas os dados.”

(The Economist, 2017)

RESUMO

Os dados são elementos cruciais para a formação de opinião, tomadas de decisões e transparência. O pleno compartilhamento de dados públicos tem fator essencial na consolidação de uma democracia mais forte. A legislação brasileira vem evoluindo para melhorar os padrões de disponibilização dos dados públicos. Este trabalho endossou a necessidade de termos dados públicos conectados,padronizados, interligados, otimizados e não somente disponibilizados em sua forma bruta. Realizando a extração de dados de diferentes fontes, em distintos formatos, e transformando para dados relacionados em uma única fonte, utilizando os processos da metodologia Knowledge Discovery in Databases (KDD), o presente trabalho demonstrou como é possível disponibilizar dados públicos de forma mais unificada, facilitando análises que anteriormente não eram possíveis ou dificultadas, em razão da falta de padronização e conexão entre os dados de diferentes bases.

Palavras-chave: Dados Públicos. KDD. Análise de dados. Banco de Dados.

ABSTRACT

Data are crucial elements for opinion formation, decision making and transparency. The full sharing of public data is an essential factor in the consolidation of a stronger democracy. Brazilian legislation has evolved to improve standards for making public data available. This work endorsed the need to have public data connected, standardized, interconnected, optimized and not only made available in its raw form. By extracting data from different sources, in different formats, and transforming it to related data in a single source, using the processes of the Knowledge Discovery in Databases (KDD) methodology, the present work demonstrated how it is possible to make public data available more efficiently, facilitating analyzes that were previously not possible or difficult.

Keywords: Public Data. KDD. Data Analysis. Databases.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interfaces*
CSV	*Comma-Separeted Values*
ETL	*Extract, Transform, Load*
IBGE	Instituto Brasileiro de Geografia e Estatísticas
JSON	*JavaScript Object Notation*
PDF	*Portable Document Format*
PIB	Produto Interno Bruto
RDF	*Resource Description Framework*
REST	*Representational State Transfer*
SQL	*Structured Query Language*
SSPDS	Secretaria de Segurança Pública e Defesa Social
URI	*Uniform Resource Identifier*
W3C	*World Wide Web Consortium*

LISTA DE SÍMBOLOS

p	Coeficiente de Correlação de Pearson

SUMÁRIO

1 INTRODUÇÃO

A Lei nº 12.527, sancionada em 18 de novembro de 2011, regulamenta o direito constitucional dos cidadãos de acesso às informações públicas, sendo aplicável aos três poderes: da União, dos estados, do Distrito Federal e dos municípios (BRASIL, 2011).

As entidades governamentais são obrigadas a disponibilizar os dados para acesso, de forma a dar transparência nos processos da administração pública e tornar possível para qualquer cidadão, entidade ou órgão, coletar, processar e analisar os dados disponibilizados. É importante salientar que não só os dados são essenciais, mas também o meio e a forma pelos quais eles são disponibilizados. Quanto mais fácil o acesso, utilização e maior facilidade de integração dos dados, maior será a contribuição deles para a tomada de decisão das autoridades governamentais e aumento da transparência (JANNUZZI, 2018).

No contexto da pandemia de Covid-19, foi possível notar o quão imprescindível é a disponibilização de dados, bem como a importância do fácil acesso, consulta e integração. A plataforma INTEGRASUS[1], disponibilizada pelo Governo do Estado do Ceará, é exemplo disso. Setores da imprensa e sociedade civil têm acesso aos dados relativos à pandemia de forma mais ágil e segura, podendo assim comunicar e conscientizar de forma mais rápida, assim como as decisões das autoridades tornam-se mais bem embasadas diante da apresentação dos dados.

As *Application Programming Interfaces* (APIs), segundo a empresa Red Hat, em artigo publicado em seu site oficial[2], são “um conjunto de definições e protocolos usado no desenvolvimento e na integração de software de aplicações”. As APIs têm papel essencial dentro do processo de integração e consulta de dados entre ambientes no contexto de compartilhamento de dados entre entidades. Há muitos repositórios de dados abertos, como os escolhidos para o fim deste trabalho, que não disponibilizam de uma API para facilitar o acesso e análise de dados neles contidos de forma filtrável e rápida.

O Governo Federal, por meio da plataforma OPENDATASUS[3], disponibiliza os dados de aplicações de vacina contra a Covid-19 através de uma API para livre

[1] INTEGRASUS. Disponível em <https://integrasus.saude.ce.gov.br/>. Acesso em 04 jul. 2022.
[2] Red Hat. APIs. Disponível em <https://www.redhat.com/pt-br/topics/api>. Acesso em 04 jul. 2022.
[3] OPENDATASUS. Disponível em <https://opendatasus.saude.gov.br/>. Acesso em 04 jul. 2022.

consulta, sendo possível obter acesso aos registros de todas as doses aplicadas em território nacional. Contudo, diante do imenso número de registros, para a completa análise, o usuário desenvolvedor que consumiria o recurso precisaria de passar por milhares de páginas, sem qualquer possibilidade de filtro e comprometimento de disponibilidade por parte da API[4].

O Governo Estadual do Ceará, através da Secretaria de Segurança Pública e Defesa Social, disponibiliza os dados relativos a crimes violentos letais e intencionais (CVLI) através de relatórios, mês a mês, com todas as ocorrências divididas por região, no formato de tabela. Neste caso, não há possibilidade de consulta dinâmica, com fins de análise secundária dos dados informados.

Dito isso, analisando as ferramentas de acesso a dados disponibilizadas pelo Governo Federal e Governo Estadual Do Ceará, é possível notar a inexistência de consulta e integração de dados relacionados à saúde e segurança pública dentro das plataformas de acesso à dados de ambos os domínios. Os únicos recursos disponíveis para acesso a estes dados, sejam através de APIs ou de planilhas em formato *Comma-Separeted Values* (CSV), trazem uma quantidade muito grande de dados, o que torna a consulta onerosa e sua análise dificultada.

Nessa perspectiva, diante dessa problemática identificada, foi possível denotar a necessidade da construção de um mecanismo capaz de extrair os dados dos repositórios. Além disso, disponibilizá-los de forma unificada e conectada para tornar viável a construção de sistemas secundários que usem desse mecanismo para fazer integração e análise visando tomadas de decisões.

Por conseguinte, questiona-se: *Como disponibilizar as informações presentes de vacinação, segurança pública e desenvolvimento social de uma forma mais otimizada e de fácil consulta? Além disso, como facilitar os sistemas secundários de secretarias de saúde estaduais a consultarem estes dados?*

Dessa maneira, o objetivo geral do trabalho é construir um mecanismo que disponibilize a integração e acesso aos dados públicos para melhoria de políticas de saúde e segurança pública.

Em razão disso, foram delimitados os seguintes objetivos específicos: (a) propor uma maneira de extração dos dados de vacinação contra a Covid-19, segurança

[4] Campanha Nacional de Vacinação contra Covid-19 - Portal Brasileiro de Dados Abertos. Disponível em:<https://dados.gov.br/dataset/covid-19-vacinacao>. Acesso em: 04 jul. 2022.

pública e desenvolvimento social, além da disponibilização via API; (b) interpretar e integrar os dados coletados a fim de investigar e extrair conhecimento; (c) investigar a relação entre os dados por meio de consultas, as quais poderiam ser propostas por estudantes ou desenvolvedores de novas soluções em saúde, por exemplo.

A seguir, no capítulo 2, será apresentado o referencial teórico utilizado para o desenvolvimento do trabalho. No capítulo 3, estão os trabalhos relacionados, que abordam problemas semelhantes aos apontados neste trabalho. No capítulo 4, será apresentada a metodologia aplicada junto ao desenvolvimento do estudo de caso em específico apresentado. Os resultados obtidos serão apresentados no capítulo 5, juntamente com as análises que se tornaram possíveis através da ferramenta. No capítulo 6, está apresentada a conclusão, em conjunto com as considerações relacionadas a trabalhos futuros.

2 REFERENCIAL TEÓRICO

O presente capítulo contará com os conceitos relacionados ao uso de dados abertos e sua relevância para a construção de sistemas integrados com informações de interesse público, contribuindo para a transparência em órgãos e instituições de caráter público. Além disso, será discorrido sobre a implementação de soluções que permitem a centralização e reuso de dados abertos, desde a extração de dados advindos de diversas fontes com a técnica de *Extract, Transform, Load* (ETL), e o seu gerenciamento com a controladoria das APIs geradas. Bem como, aplicando sobre esses dados, as etapas de descoberta de conhecimento, ou em inglês, *Knowledge Discovery in Databases* (KDD).

2.1 Dados Abertos

Desde a regulamentação da lei de acesso à informação, tornou-se cada vez mais necessário a multiplicação de meios com o fim de disponibilizar os dados públicos a fim de os tornarem abertos. De acordo com a *Open Knowledge Foundation* (OPEN KNOWLEDGE FOUNDATION, 2021), dados abertos são aqueles livres para uso, modificação e compartilhamento, sem qualquer restrição.

Em dezembro de 2007, um grupo de trabalho com 30 pessoas, denominado Open Data Gov[5], delimitou os 8 princípios base para os dados governamentais abertos, sendo eles: completo, sem limitações; primário, coletados em sua origem; oportuno, disponibilizados o mais rápido possível; Acessível, disponíveis para todos; processável por máquina; não discriminatório; não proprietário; e por fim, livre de licenças.

A partir destes princípios, David Eaves delimitou 3 leis dos dados abertos: se não pode ser facilmente achado, não existe; se não estiver disponível em formatos que possam ser compreendidos por máquinas, eles não podem ser aproveitados; se por algum motivo uma legislação impede o seu compartilhamento, ele perde a sua utilidade (EAVES, 2009).

Além disso, em 2011 foi criado a , aliança global de países “para promover uma forma de governança mais transparente, participativa, inclusiva e responsável” (OPEN GOVERNMENT PARTNERSHIP, 2011). Os países participantes asseguram

[5] Open Data Gov. Disponível em <https://opengovdata.org/> Acesso em 4 jul. 2022

o compromisso de promover o aumento da disponibilidade de dados relativos às atividades exercidas pelo poder público.

Entre as iniciativas para o uso de dados abertos no Brasil, foi criado o Portal Brasileiro de Dados Abertos[6], ferramenta desenvolvida e disponibilizada pelo governo para facilitar o acesso a informações públicas, auxiliando o trabalho de quem deseja consumir os dados ou transformá-los a fim de buscar novas informações ou simplesmente por motivos de apresentação. Iniciativas como esta são essenciais para que a flexibilidade dos dados tornem-se possíveis, visto que há a necessidade de se criar ecossistemas novos a partir dos dados disponibilizados, como APIs, serviços e bases de dados.

Dito isto, é importante salientar as ferramentas que podem auxiliar no processo de desenvolvimento de aplicações com o uso de dados abertos. Dentro do Portal Brasileiro de Dados Abertos, há o chamado Kit para dados abertos, um conjunto de documentos e ferramentas para a implementação do uso de dados públicos em uma instituição. Uma ferramenta bastante utilizada é a CSV to API[7], que gera APIs dinamicamente a partir de arquivos CSV.

2.2 Gerenciamento De Dados Abertos

Soluções a partir de dados abertos estão sendo propostas e uma tendência que vem chamando a atenção é a construção de Big Data[8] para auxiliar nas políticas públicas a partir do conglomerado de dados colhidos. Os dados massivos tem grande potencial para aumentar a eficiência da administração pública no contexto do desenvolvimento urbano (CERDEIRA *et al.*, 2020)).

No contexto da pandemia de Covid-19, o uso de dados públicos se tornou essencial no controle e conscientização da população com relação ao vírus. Como explicado por Jeni Tennison , em entrevista à Forbes, “Tornar os dados abertos, através da publicação na internet, em planilhas, sem restrições de acesso, é a melhor maneira de garantir que eles possam ser usados pelas pessoas que mais precisam.” (FORBES, 2020)

---

6 Portal Brasileiro de Dados Abertos. Disponível em <https://dados.gov.br/>. Acesso em 4 jul 2022.

7 CSV to API. Disponível em <https://github.com/project-open-data/csv-to-api>. Acesso em 4 jul 2022.

8 O Que é Big Data? | Oracle Brasil. Disponível em: <https://www.oracle.com/br/big-data/what-is-big-data/>. Acesso em 4 jul 2022 .

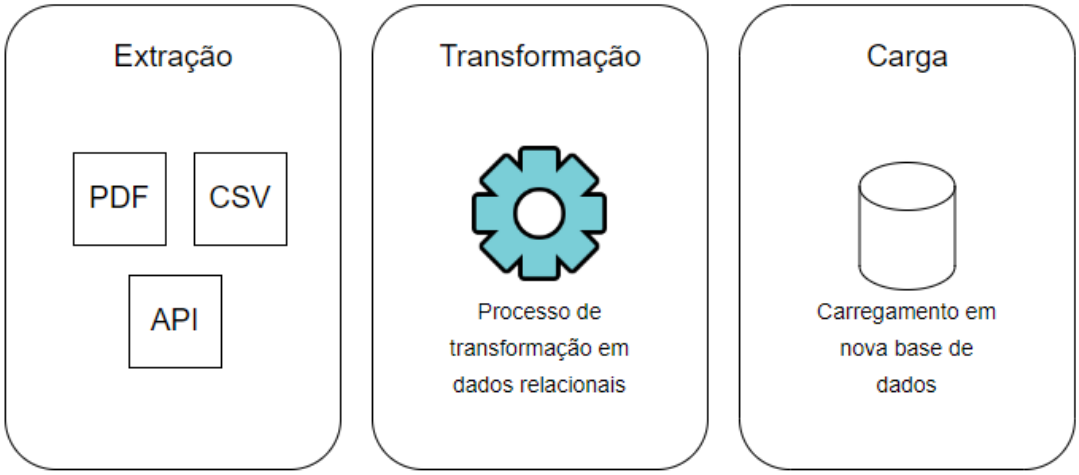
Com a grande quantidade de dados públicos disponíveis para consulta, advindos de várias fontes e com diversos formatos diferentes, surge a necessidade de agrupar dados e administrar as informações por meio de barramentos de serviços ou gerenciamento das chamadas com diversas origens.

Para isso, é preciso abordar o termo Linked Data (dados conectados). O *World Wide Web Consortium* (W3C), consórcio mundial para tratamento de padrões na web, define o termo como os conjuntos e coleções de dados relacionados à Web. É possível usar esse padrão para traçar a ligação entre dados abertos do poder público, dessa forma, é possível criar um *Linked Data* a partir dos dados coletados por este trabalho (W3C, 2011).

2.3 Ferramentas e Técnicas Para Manipulação De Dados

Para tornar possível a integração de sistemas como o que se deseja fazer neste projeto com os dados de vacinação, dados de homicídio da segurança pública e desenvolvimento social, existem técnicas amplamente utilizadas. ETL (conforme apresentado na Figura 1) é uma técnica utilizada para extrair dados de diversas fontes, junto de otimização e classificação para a inserção em uma nova base de dados (FERREIRA *et al.*, 2010).

Figura 1 – Desenho do processo de ETL



Fonte: Elaborado pelo autor (2022).

Dentro do escopo deste projeto de pesquisa, há necessidade da extração dos dados advindos de diferentes origens dos sistemas estaduais e federais, tornando

necessário o uso de técnicas de extração dos dados já presentes, sua transformação em dados relacionais em banco de dados, junto de otimização, homogeneização e classificação, além da devida carga dentro da estrutura previamente montada para receber os dados oriundos das diversas fontes selecionadas no primeiro processo de extração, para a formação de uma nova base.

Para disponibilizar os recursos oriundos de várias bases através de uma plataforma centralizada de integração e reuso de dados compartilhados, é comumente utilizada a técnica de construção de APIs que criam uma nova interface com a base de dados criada a partir dessa nova junção e classificação de dados.

As bases selecionadas dispõem de dados extremamente relevantes e que podem ser heterogêneos. Visto isso, surge a necessidade de implementar um mecanismo de reuso e centralização desses dados, os tornando disponíveis através de serviços em que possam ser consultados separadamente via APIs, mas oriundos de uma única fonte.

2.4 O Processo De KDD e As Ferramentas Utilizadas Nas Etapas

Para o desenvolvimento deste projeto foram utilizadas as etapas definidas no *Knowledge Discovery in Databases* (KDD), que, segundo (ADRIAANS, 1996), permite o descobrimento de conhecimento útil previamente desconhecido a partir de uma base de dados. Portanto, utilizando-se das bases selecionadas, foi possível gerar conhecimento integrado a partir dos dados unificados. O processo de KDD, segundo (FAYYAD *et al.*, 1996), pode ser dividido nas seguintes etapas:

- **Seleção**: A escolha das bases de dados a serem utilizadas no processo, definindo todo o conjunto.
- **Pré-Processamento**: Ocorre a primeira análise e limpeza primária de dados em duplicidade, com redundância, e dados identificados que claramente não fazem parte daquele conjunto ou contém algum tipo de ruído.
- **Transformação**: Nesta etapa ocorre a transformação dos dados previamente processados, formatados e que serão armazenados na nova base, com o fim de disponibilizar o acesso dos motores que farão a seguinte análise.

- **Data Mining ou simplesmente Análise de Dados**: Esta é a etapa mais importante dentro de todo o processo do KDD. É a partir daqui que é possível de fato ser gerado o conhecimento útil previamente desconhecido. São aplicados algoritmos com o fim de analisar os dados, gerando assim padrões.
- **Interpretação**: Esta etapa se destina a interpretar o conhecimento gerado a partir de padrões da etapa anterior, havendo asserções de relevância ou não, e permitindo a resolução de questões de análise previamente especificadas no processo.

A Figura 2 ilustra o processo de KDD. A seguir, serão discutidas algumas ferramentas e tecnologias utilizadas neste trabalho nas etapas do processo de KDD.

Figura 2 – Desenho do processo de KDD

Fonte: (Fayyad, apud GONÇALVES, 2001)

Inicialmente, uma das tecnologias utilizadas neste trabalho foi a linguagem *Structured Query Language* (SQL). Criada pelos pesquisadores Raymond Boyce e Donald Chamberlin, no intuito de criar uma linguagem padrão para acesso e manipulação de bancos de dados relacionais, a *Structured Query Language* (SQL) (CHAMBERLIN; BOYCE, 1974) é uma linguagem declarativa que se tornou amplamente utilizada na indústria e academia. Suas vantagens estão desde a sintaxe simplificada, fácil uso, entendimento e a capacidade de processamento de quantidade elevada de dados.

Outra tecnologia utilizada foi **GraphQL**. Essencialmente, é uma linguagem de consulta a dados, criada pela empresa Facebook em 2016 para o desenvolvimento de APIs como forma alternativa a outras já existentes. Sua notação se assemelha bastante aos objetos *JavaScript Object Notation* (JSON), amplamente utilizados no mercado. Por mais que seja muitas vezes utilizada como uma linguagem de banco de dados, não se encaixa nesta categoria. Sua ideia principal é trabalhar como um interpretador e mapear os campos da base de dados com seus respectivos tipos de

retorno necessários através de um esquema (FACEBOOK, 2016).

Dentro do processo de desenvolvimento de software, há uma busca por padrões a serem seguidos com o intuito de uniformizar a produção de produtos computacionais. O padrão de **arquitetura *Representational State Transfer* (REST)**, foi desenvolvido por Roy Fielding com o intuito de padronizar o desenvolvimento de APIs. Este modelo representa um conjunto de características específicas, como a representação de estado e sua transferência a partir de protocolos na web, através de identificadores únicos, denominados *Uniform Resource Identifier* (URI) (FIELDING *et al.*, 2017).

O software Hasura disponibiliza um serviço de conexão, manipulação e processamento de bases de dados, gerando API GraphQL ou Rest através da ferramenta. Este software atua como BaaS (*Backend as a service*), disponibilizando recursos de processamento de regras de negócio, autenticação e gerência de rotas de uma aplicação. A grande vantagem da utilização da ferramenta se dá pela otimização de tempo de desenvolvimento proporcionado por ela, visto que é possível criar consultas a uma base de dados e geração de uma API rapidamente, principalmente quando o grande objetivo é disponibilizar acesso a dados contidos em uma base. O software é instalável e executável a partir de uma imagem de **contêiner**, o que facilita sua replicação e uso (HASURA, 2018).

A **virtualização de ambientes**, técnica que permite a criação de múltiplas instâncias de sistema operacional dentro de uma máquina, é amplamente utilizada para dar independência entre aplicações e facilidade de manutenção. A **conteinerização**, técnica que virtualiza recursos a nível de sistema operacional, possibilita o uso de um mesmo hospedeiro físico por múltiplos usuários ou aplicações. Se comparada, por exemplo, com virtualização baseada em *hypervisor*, contêineres destacam-se na maioria dos aspectos(SILVA, 2017).

Por exemplo, contêineres são mais eficientes, rápidos e compactos, utilizando-se somente de dependências necessárias para a execução da aplicação. Contêineres trabalham com múltiplas tarefas, e quando se depara com falhas, pode ser facilmente recriado ou reinicializado, sendo bastante resiliente (FERNANDEZ *et al.*, 2017).

3 TRABALHOS RELACIONADOS

O seguinte capítulo deste trabalho tem como finalidade abordar trabalhos relacionados aos objetivos previamente introduzidos. Os trabalhos foram encontrados através de consulta pelas palavras chave deste trabalho no Google Acadêmico. Foram selecionados devido a busca comum por melhora na disponibilização e acesso a dados públicos de grande relevância ou interesse.

3.1 Avaliação Da Qualidade De Dados Públicos De Extração E Tratamento De Produtos Minerais

Este trabalho foi apresentado no Simpósio de Engenharia de Produção da Universidade de Catalão, com o intuito de avaliar a qualidade dos dados públicos referentes à extração e tratamento de produtos minerais disponibilizados pelo Portal Brasileiro de Dados Abertos (CORDEIRO, 2021).

Foram realizadas verificações exploratórias sobre a base de dados, visando delimitar os dados constantes e detectar a presença de dados ausentes ou inconsistentes dentro do conjunto de cada atributo contido nos dados. Também foi feita uma avaliação sobre os campos numéricos da base de dados com o fim de identificar a presença de dados fora do padrão estabelecido no conjunto, também chamados de *outliers*.

Como resultados obtidos, os autores detectaram a ocorrência de não homogeneidade dentro do conjunto de dados em atributos considerados extremamente relevantes para a identificação de cada entidade na base. Também foi possível aferir a quantidade relevante de dados ausentes dentro do conjunto.

3.2 Web Scraping Em Dados Públicos: Método Para Extração De Dados Dos Gastos Públicos Dos Vereadores Da Câmara Municipal De Belo Horizonte

Este trabalho foi desenvolvido com o objetivo de demonstrar uma forma de extrair dados de bases públicas e disponibilizá-los através de dados abertos. Para o estudo de caso foi escolhido os dados relativos aos gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte, obtidos através do portal da transparência do município (ASSIS; GOMIDE, 2021).

A metodologia utilizada se baseou em uma pesquisa bibliográfica em dados abertos, com fins de mapear os dados a serem objeto de extração e delimitar os pontos de acesso para a obtenção destes dados. Foi realizada a análise de conteúdo do site que hospeda as informações, seguida da utilização de bibliotecas *Selenium* e *Pandas* da linguagem *python* com o intuito de automatizar extração, manipulação e transformação dos dados. Através de elementos da página web, foi realizado o mapeamento dos dados não estruturados para uma estrutura armazenável.

A partir dos dados obtidos e estruturados, foi possível realizar análises sob novas perspectivas, agregações e granularidade. Foi detectada uma deficiência na forma com que os dados estavam anteriormente disponibilizados e verificada a grande necessidade de se otimizar a maneira em que os dados públicos são compartilhados.

3.3 Publicando Dados Na Web De Dados: Um Relato De Experiência Na Automatização Dos Processos De Extração, Transformação E Carga De Dados Abertos Provenientes Do Portal Dados.gov.br

Este trabalho foi apresentado no 18º encontro nacional de pesquisa em ciência da informação, com o objetivo de demonstrar a experiência obtida durante processos de extração e carga de dados abertos oriundos do portal brasileiro de dados abertos. Também foi objetivo criar elos de ligação entre os dados, os transformando em *Linked Data*, aumentando também a classificação de qualidade dos dados abertos disponibilizados (RAUTENBERG *et al.*, 2017).

Os autores utilizaram dados relativos ao Programa Nacional de Fortalecimento da Agricultura Familiar para estabelecer um *workflow* de tratamento, manipulação e processamento dos dados. Foi utilizado *scripts* para a conversão de formatos de dados, obtendo arquivos CSV e posteriormente processando estes arquivos com o fim de transformá-los em arquivo do formato *Resource Description Framework* (RDF). Os dados foram carregados em uma base de dados, e consequentemente puderam ser objeto de análise.

Diante da carga realizada, foram realizadas análises exploratórias que permitiram a identificação e classificação de contratos realizados pela entidade, listando municípios e outras informações de extrema relevância, demonstrando a finalidade do produto criado.

3.4 Identificação De Mecanismos Para A Ampliação Da Transparência Em Portais De Dados Abertos: Uma Análise No Contexto Brasileiro

Este trabalho foi desenvolvido dentro do contexto da necessidade de melhoria na forma com que os dados abertos são disponibilizados. Vários são os padrões e exigências para se inferir o que são dados compartilhados de forma satisfatória, mas nem sempre é possível identificar isso na forma com que os portais de dados abertos funcionam. Visto isso, surgiu-se a necessidade de identificar mecanismos para a ampliação da transparência em portais de dados abertos no contexto brasileiro (KLEIN *et al.*, 2018).

A pesquisa adquiriu um caráter exploratório com o fim de identificar todo o cenário da disponibilização de dados abertos no contexto brasileiro. Em cada uma das 5 fases da pesquisa, os pesquisadores utilizaram ferramentas para aferir a qualidade das informações disponibilizadas à época, se elas seguiam os padrões já estabelecidos para o respectivo tipo de informação compartilhada e o cálculo do nível de transparência dos dados públicos disponibilizados.

O trabalho obteve resultados ao conseguir classificar e definir o nível de transparência dos dados a partir de sua fonte. Segundo os autores, todos os mecanismos para a ampliação de transparência de dados públicos se mostraram relevantes ou extremamente relevantes para pessoas que utilizam dados abertos. Como destaque, fica-se concluído que, em termos de transparência governamental, não basta somente a informação ser adequada ao propósito de estar ali disponível, mas estar adequada ao uso daqueles que irão consumi-la.

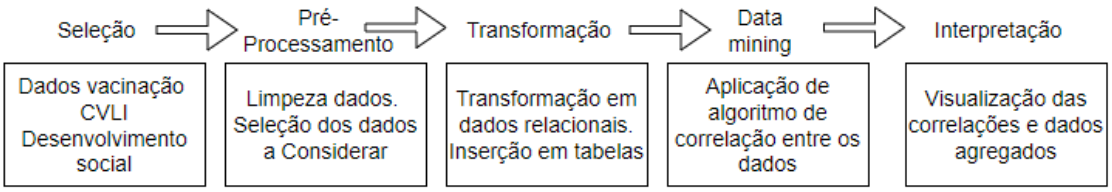
Analisando os trabalhos relacionados, é possível traçar paralelos com o presente trabalho. Nenhum deles construiu uma API para dar possibilidade de consumo aos dados extraídos, pois houve foco em exploração, extração de dados, e avaliação da qualidade de dados públicos disponibilizados. As fontes de dados escolhidas neste trabalho, em conjunto, não foram objetos de estudo nos trabalhos relacionados.

4 METODOLOGIA / DESENVOLVIMENTO

A busca e necessidade pela disponibilização de dados abertos de forma clara e de fácil interação tem crescido exponencialmente ao longo dos últimos anos. Como foi explicitado anteriormente, entidades governamentais de todo o mundo têm feito acordos para que os dados possam ser melhor consumidos pelo seu público final. A construção de ferramentas que contribuam para o melhor acesso a dados abertos, seja pela sociedade civil ou pelo poder público, constitui relevante papel do projeto na garantia do acesso democrático e eficiente para todos, seja na elucidação de uma dúvida ao cidadão, ou na tomada de decisões de autoridades.

O processo de KDD foi aplicado a fim de garantir que as informações desconhecidas presentes nas bases de dados sejam melhor visualizadas, com otimização e junção de diferentes visões, conforme apresentado na Figura 3.

Figura 3 – Fluxo do KDD aplicado neste trabalho



Fonte: Elaborado pelo autor (2022).

4.1 Seleção Dos Dados

A definição do conjunto de dados foi baseada na grande relevância que as informações neles obtidas tem para a sociedade em geral e para a tomada de decisões do poder público. Diante disso, o conjunto escolhido foi:

- **Dados de Vacinação contra a Covid-19 do Estado do Ceará**: A plataforma OPENDATASUS disponibiliza por meio de uma *Application Programming Interfaces* (API) em que podem ser consultados todos os registros de vacinação contra a Covid-19 no território nacional. Contudo, com o fim de melhorar a análise e granularidade dos dados, foi escolhido capturar somente os dados de aplicações de doses no Estado do Ceará

no ano de 2021. Em razão desta escolha deste intervalo temporal, foi necessário limitar também os registros de aplicações de dose. Dessa forma, os dados constam apenas as doses relativas à primeira e segunda doses, além da dose única.

- **Dados de Crimes Violentos Letais e Intencionais no Estado do Ceará**: A Secretaria de Segurança Pública do Estado do Ceará disponibiliza os registros de CVLI em arquivos no formato *Portable Document Format* (PDF), contendo tabelas que mostram os casos individuais, de acordo com parâmetros de região e características do fato. Foram selecionados os documentos que continham dados de todos os registros de CVLI do ano de 2021.
- **Dados de Desenvolvimento Social dos municípios do Estado do Ceará**: Com o fim de dar mais opções de análise e junção dos dados relacionados aos municípios do Estado, foi selecionado os dados de desenvolvimento social de cada município do Estado, com origem de planilha retirada do Instituto Brasileiro de Geografia e Estatísticas (IBGE).

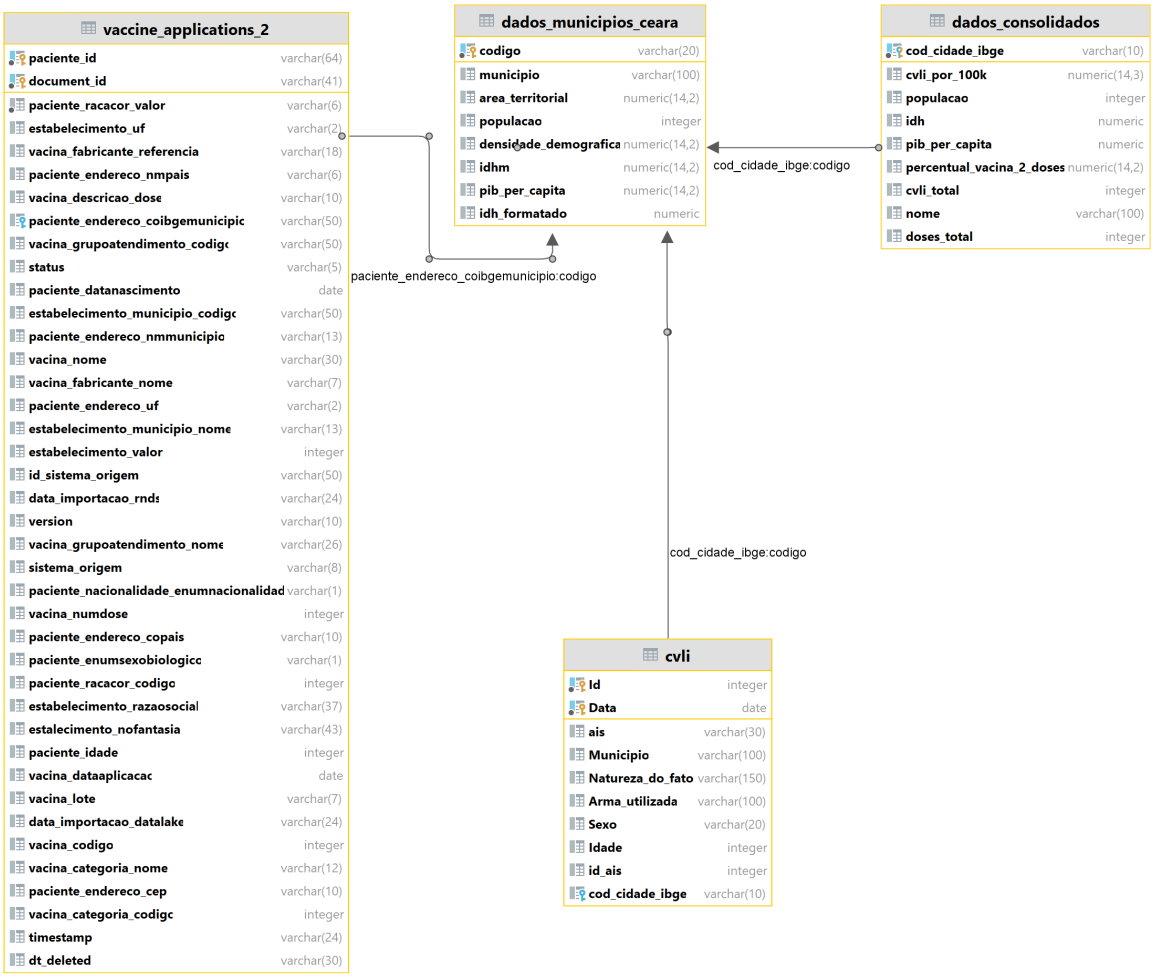
O objetivo é integrar essas bases e descobrir conhecimento a partir delas. Além de oferecer, uma forma de publicação desses dados de maneira útil, facilitando que essas informações sejam consultadas.

4.2 Preparação De Ambiente

Inicialmente, visando receber dados de diversas fontes, foi arquitetada uma estrutura de tabelas em banco de dados relacional, utilizando-se de linguagem SQL e o sistema gerenciador de banco de dados PostgreSQL. A estrutura necessária foi concebida mediante análise dos dados que se deseja armazenar.

Como se pode observar a partir do diagrama na Figura 4, foram criadas três tabelas base com o intuito de armazenar os dados extraídos das distintas fontes selecionadas. Uma quarta tabela, de dados consolidados, foi criada com o intuito de armazenar os dados já agrupados, processados e com informações de extrema importância para futuras análises. Com isto, como a base de dados atual é estática, será poupado tempo de processamento com agregação e seleção de dados, otimizando as consultas.

Figura 4 – Diagrama UML da estrutura de banco de dados utilizada no projeto



Fonte: Elaborado pelo autor (2022).

4.3 Pré Processamento

Diante de uma análise inicial criteriosa, foram identificadas informações em duplicidade dentro das bases de dados selecionadas. A base de vacinação contra a Covid-19 continham dados ruidosos, então após a obtenção dos arquivos em extensão CSV com todos os registros de vacinação do Estado, foi feita uma limpeza inicial, com a retirada de dados nulos.

No caso dos dados de CVLI, os registros encontravam-se em tabelas dentro de um arquivo PDF, tornando necessária a sua extração automatizada. Foi realizado o *download* de todos os 12 arquivos referentes ao espaço temporal selecionado, um para cada mês. A biblioteca de software Tabula[9] necessita que as das tabelas analisadas sejam delimitadas por bordas, com o fim de separar o conteúdo de cada uma. Dito isto,

[9] Tabula Python. Disponível em <https://pypi.org/project/tabula-py/>. Acesso em 4 jul. 2022.

foi realizada uma transformação no arquivo pdf, alterando a tabela para inserir todas as bordas de cada tabela.

Os dados de desenvolvimento social foram os que exigiram menos processamento prévio, em razão de já estarem bem delimitados e com informações precisas relativas a cada município do Estado do Ceará.

4.4 Transformação

Os dados previamente selecionados foram transformados com a finalidade de adquirirem o mesmo formato, neste caso o de dados relacionais em banco de dados SQL. As tabelas foram previamente geradas com a intenção de receber a base de informações sem perda de dados relevantes. Os dados de vacinação contra a Covid-19 foram extraídos em formato CSV e carregados diretamente no banco de dados, através da ferramenta Datagrip[10], onde há uma importação através de mapeamento entre os campos presentes no arquivo para os respectivos campos da tabela no banco de dados.

Os dados de CVLI, que encontravam-se em tabelas no formato PDF (veja na Figura 5) disponíveis para download na plataforma da Secretaria de Segurança Pública e Defesa Social (SSPDS), divididos por mês, foram reunidos em um único diretório, e a partir de um *script* na linguagem de programação *Python* (veja na Figura 6), utilizando-se das bibliotecas *Tabula* e *Pandas*[11]. Por meio dessas tecnologias, foi possível extrair as tabelas dos documentos PDF, transformando-as em arquivos CSV, prontos para serem importados ao banco de dados pela mesma via em que os dados de vacinação contra a Covid-19 foram.

A biblioteca Tabula lê o arquivo PDF e identifica as tabelas, separando e agrupando as células, transformando-as em Dataframes, estruturas da biblioteca Pandas, que podem ser manipuladas e transformadas em outros tipos de dados, como arquivos CSV.

Os dados relativos ao desenvolvimento social dos municípios foram extraídos da plataforma oficial do Instituto Brasileiro de Geografia e Estatísticas (IBGE) já em formato CSV, prontos para a importação no banco de dados previamente criado. A

[10] Datagrip, Disponível em <https://www.jetbrains.com/datagrip/>. Acesso em 4 jul. 2022
[11] Pandas, Disponível em <https://pandas.pydata.org/>. Acesso em 4 jul. 2022

Figura 5 – Exemplo de arquivo com dados de CVLI disponibilizado

ID	AIS	MUNICÍPIO	NATUREZA DO FATO	ARMA-UTILZADA	DATA	SEXO	IDADE
1	AIS 11	Paraipaba	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Feminino	26
2	AIS 12	Pacatuba	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	32
3	AIS 2	Fortaleza	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	30
4	AIS 9	Fortaleza	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	30
5	AIS 12	Itaitinga	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	50
6	AIS 18	Aracati	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	17
7	AIS 19	Juazeiro do Norte	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	26
8	AIS 21	Iguatu	HOMICIDIO DOLOSO	Arma branca	01/01/2021	Masculino	16
9	AIS 14	Sobral	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	26
10	AIS 22	Mombaça	LESAO CORPORAL SEGUIDA DE MORTE	Outros meios	01/01/2021	Feminino	15
11	AIS 15	Canindé	HOMICIDIO DOLOSO	Arma de fogo	01/01/2021	Masculino	-
12	AIS 9	Fortaleza	ROUBO SEGUIDO DE MORTE (LATROCINIO)	Arma de fogo	01/01/2021	Masculino	20
13	AIS 11	São Gonçalo do Amarante	HOMICIDIO DOLOSO	Arma branca	02/01/2021	Masculino	29
14	AIS 19	Brejo Santo	HOMICIDIO DOLOSO	Arma de fogo	02/01/2021	Masculino	42

Fonte: Elaborado pelo autor (2022).

Figura 6 – Trecho de código que exemplifica a extração de dados oriundos de arquivo PDF

```
df1 = read_pdf("CVLI-1.pdf", encoding='utf-8', pages='all',multiple_tables=False,lattice=True)[0]

print(df1)
```

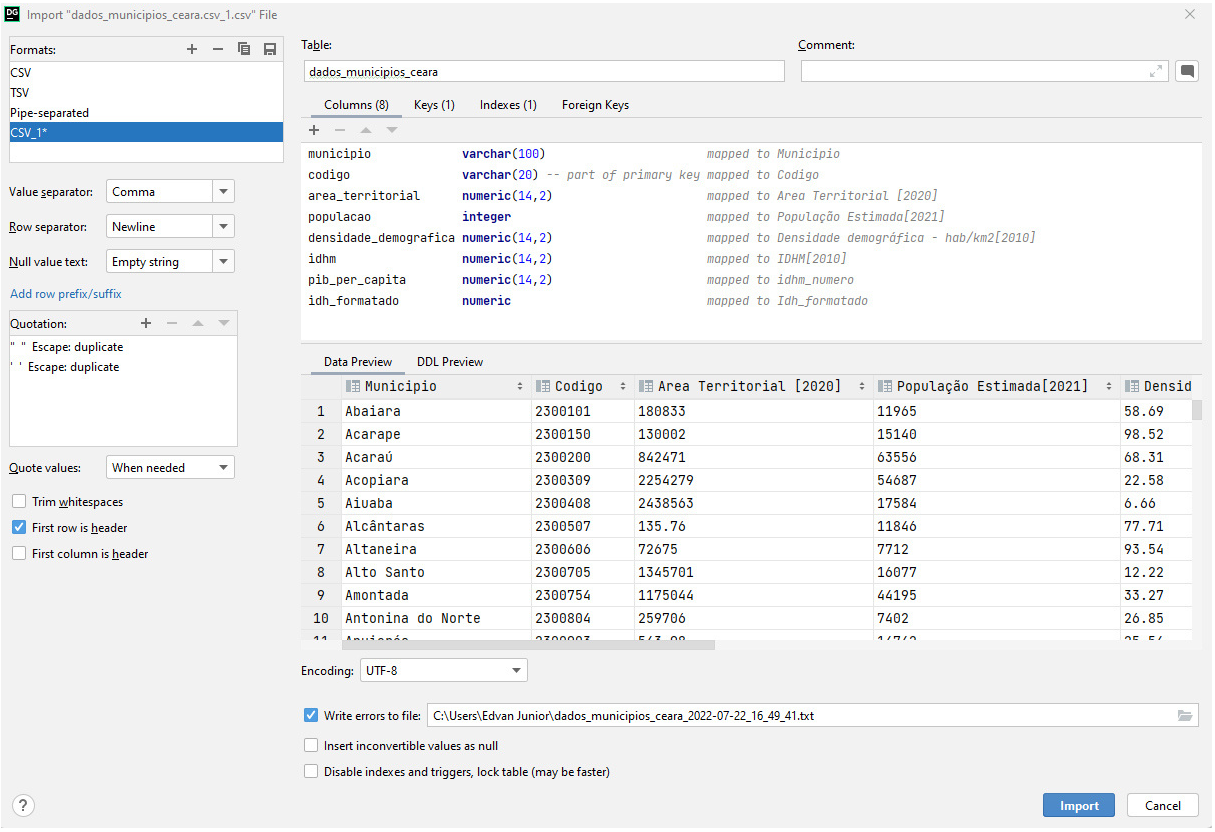
```
pdftocsv
"C:\Users\Edvan Junior\testes\venv\Scripts\python.exe" "C:/Users/Edvan Junior/testes/pdftocsv.py"
      ID     AIS          MUNICÍPIO  ...        DATA       SEXO IDADE
0      1  AIS 11          Paraipaba  ...  01/01/2021   Feminino    26
1      2  AIS 12           Pacatuba  ...  01/01/2021  Masculino    32
2      3   AIS 2          Fortaleza  ...  01/01/2021  Masculino    30
3      4   AIS 9          Fortaleza  ...  01/01/2021  Masculino    30
4      5  AIS 12          Itaitinga  ...  01/01/2021  Masculino    50
..   ...     ...                ...  ...         ...        ...   ...
302  302  AIS 12         Maranguape  ...  31/01/2021  Masculino    30
303  303   AIS 6          Fortaleza  ...  31/01/2021  Masculino    22
304  304  AIS 19             Jardim  ...  31/01/2021  Masculino    40
305  305  AIS 21             Iguatu  ...  31/01/2021  Masculino    18
306  306  AIS 19  Juazeiro do Norte  ...  31/01/2021  Masculino    39

[307 rows x 8 columns]
```

Fonte: Elaborado pelo autor (2022).

Figura 7 exemplifica a importação dos dados oriundos de arquivo CSV no banco de dados relacional.

Figura 7 – Exemplo de importação de arquivos CSV no banco de dados.



Fonte: Elaborado pelo autor (2022).

4.5 Data Mining

Uma vez com os dados importados ao banco de dados, foi preciso fazer o processo necessário para tornar possível a análise e conhecimento a partir das novas possibilidades criadas a partir da agregação de dados presentes. Foi escolhido, para fins de análise territorial, uma agregação de dados por município, com a intenção de entregar uma conhecimento útil relacionado com a situação de cada município do Estado do Ceará.

Diante disto, foi criada uma nova tabela, para receber os dados consolidados já calculados, visto que a base de dados não é dinâmica. Para inserirmos dados nesta nova tabela, foi executado um *script* em linguagem SQL agregando os dados na granularidade necessária, além de filtrar os dados de maneira a selecionar os dados a se comparar (veja na Figura 8).

Com a finalidade de gerar a capacidade de interpretação, análise e consumo dos dados advindos da base de dados gerada, foi desenvolvida uma API com o padrão REST, que pode ser utilizada para consultar os dados de forma rápida, eficaz e

Figura 8 – Script SQL usado para popularizar a tabela de dados consolidados.

```sql
insert into dados_consolidados (cod_cidade_ibge, cvli_por_100k, populacao, idh, pib_per_capita,
                                percentual_vacina_2_doses, cvli_total, nome, doses_total)
select d.codigo,
       d.cvli_rate,
       d.populacao,
       d.idh_formatado,
       d.pib_per_capita,
       cast((d.cont / cast(d.populacao as decimal)) * 100000 as decimal(14, 2)),
       d.count_cvli,
       d.municipio,
       d.cont
from (
         select d.municipio,
                d.codigo,
                count(document_id) as cont,
                d.populacao,
                cvli.rate          as cvli_rate,
                cvli.count_cvli,
                d.idh_formatado,
                d.pib_per_capita
         from vaccine_applications
                  inner join dados_municipios_ceara d on "left"(d.codigo, 6) = paciente_endereco_coibgemunicipio
                  left join (select cod_cidade_ibge,
                                    count(*)                                          as count_cvli,
                                    d.municipio,
                                    populacao,
                                    ((count(*) / cast(populacao as decimal)) * 10000) as rate
                             from cvli
                                      inner join dados_municipios_ceara d on d.codigo = cod_cidade_ibge
                             group by cod_cidade_ibge, d.municipio, populacao) cvli on cvli.cod_cidade_ibge = d.codigo
         where upper(paciente_endereco_uf) = 'ce'
           and vacina_descricao_dose in ('2ª dose', 'única')
         group by d.codigo, d.municipio, d.populacao, cvli.rate, count_cvli, d.idh_formatado, d.pib_per_capita
     ) d;
```

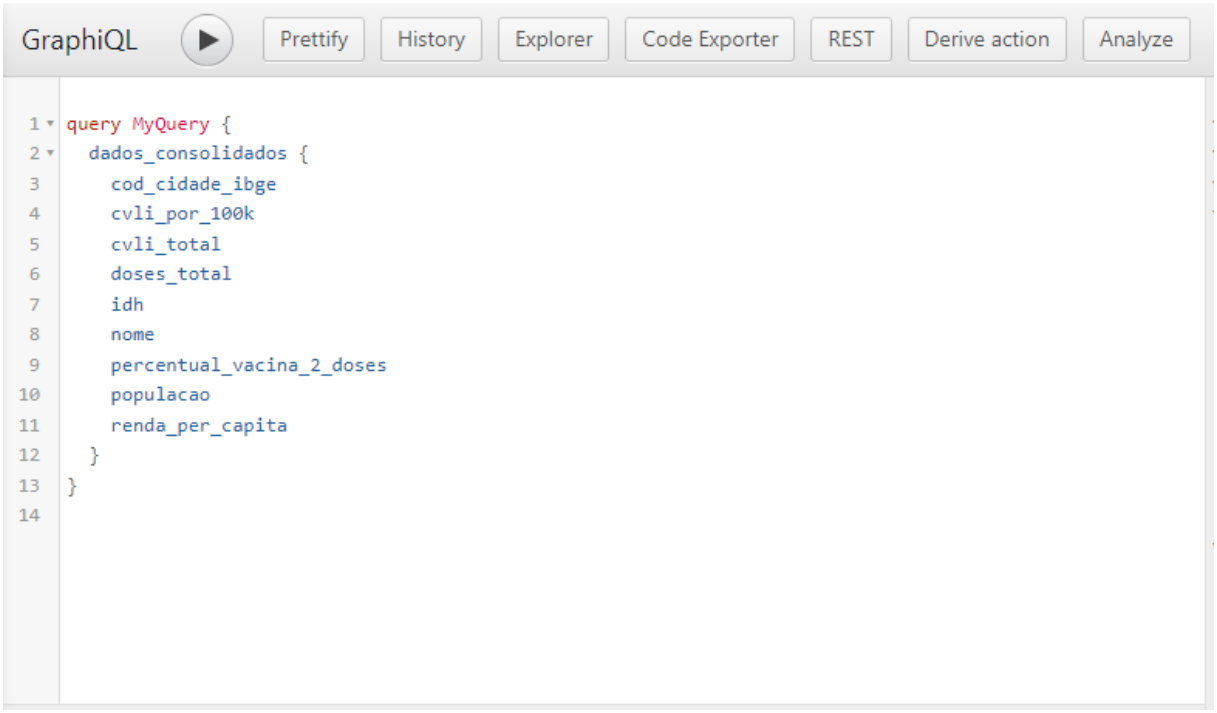
Fonte: Elaborado pelo autor (2022).

otimizada. A partir do consumo dos *endpoints*, os usuários estarão a elaborar análises de acordo com sua necessidade e ideias.

A API foi desenvolvida com a ferramenta Hasura, que facilita o desenvolvimento de API com rápida e configurável conexão a bancos de dados. Nesta ferramenta é possível gerar endpoints, que são pontos de acesso ou rotas que levam o cliente até o serviço desejado no lado do servidor. A geração é feita através de consultas GraphQL, como demonstrado na Figura 9, retornando objetos do tipo *JavaScript Object Notation* (JSON). A ferramenta Hasura é executável através da virtualização de seu ambiente por meio de contêiner, tecnologia que facilita o desenvolvimento de aplicações e captura de dependências das mesmas.

A partir do consumo destes *endpoints* gerados pela API, é possível analisar os dados e trazer conhecimento útil a partir dos mesmos.

Figura 9 – Exemplo de consulta GraphQL na através da ferramenta Hasura.



Fonte: Elaborado pelo autor (2022).

4.6 Disponibilização para o Público

Como forma de tornar a ferramenta criada pública, foi criado um repositório na ferramenta *Github*[12] com todos arquivos e instruções necessárias para a execução da ferramenta. Junto a isso, está disponibilizada toda a base de dados para migração, leitura e análise por toda e qualquer pessoa interessada no propósito deste trabalho.

[12] Repositório do Projeto. Disponível em <https://github.com/edvanjunior/publicData>. Acesso em 4 jul. 2022

5 RESULTADOS

Com o objetivo de validar a criação da ferramenta que este trabalho teve por fim, foram delimitadas questões de análise, e estas precisavam ser respondidas a partir de consulta na API construída durante este trabalho. Neste capítulo está contido o que foi obtido dentro de cada questão de análise.

As questões de análise são: (a) Qual a correlação entre a taxa de CVLI por 100 mil habitantes de um município e o percentual de população vacinada com duas doses ou dose única de um município?; (b) Qual a correlação entre a taxa de CVLI por 100 mil habitantes de um município e o Produto Interno Bruto (PIB)[13] per capita municipal?; (c) Qual a correlação entre o percentual de população vacinada com duas doses ou dose única de um município e o PIB per capita municipal?; (d) Qual a correlação entre número de habitantes de um município e a taxa de CVLI por 100 mil habitantes?; (e) Qual a correlação entre o índice de desenvolvimento humano de um município e a taxa de CVLI por 100 mil habitantes?; (f) Há cidades que aparecem tanto na lista de cidades com maiores taxas de CVLI quanto na lista de cidades com menor percentual de vacinação?. Importante salientar que este capítulo de resultados traz análises que correspondem a etapa de interpretação presente no processo KDD.

5.1 Correlação

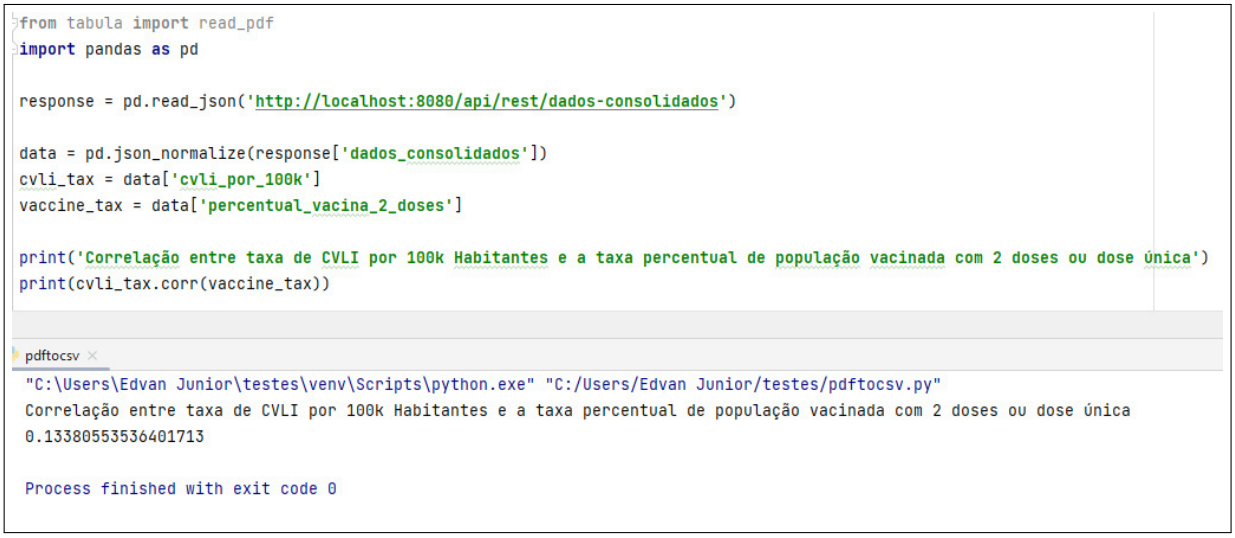
Com o objetivo de investigar a correlação entre as variáveis agregadas por município do Estado do Ceará, foi utilizada a biblioteca Pandas, amplamente usada para análise de dados. A partir do conhecimento de 2 variáveis em um conjunto, a biblioteca é capaz de aferir o coeficiente de Pearson(p)[14] e avaliar a correlação entre as variáveis.

Através do consumo da api construída neste trabalho, foi possível capturar os dados necessários para fazer a análise de correlação entre as variáveis e responder às questões previamente levantadas. Na Figura 10, o script executa um método responsável pelo cálculo de coeficiente de correlação, mediante o conhecimento das séries de dados informados.

[13] Produto Interno Bruto. Disponível em <https://ibge.gov.br/explica/pib.php/>. Acesso em 4 jul. 2022

[14] Peason Correlation. Disponível em <https://libguides.library.kent.edu/SPSS/PearsonCorr>. Acesso em 4 jul. 2022

Figura 10 – Exemplo de código usado para gerar coeficiente de pearson através da biblioteca Pandas



Fonte: Elaborada pelo autor (2022).

5.1.1 Correlação Entre Taxa De Cvli e Vacinação

A primeira questão de análise investigada foi se existe correlação entre as variáveis: taxa de CVLI por 100 mil habitantes e o percentual de população vacinada com 2 doses ou dose única no ano de 2021. O resultado está apresentado na Tabela 1. A correlação de Pearson mede a força da relação linear entre duas variáveis. Tem um valor entre -1 e 1. O valor de -1 significa uma correlação linear negativa total, 0 sendo nenhuma correlação e + 1 significando uma correlação total positiva. Pelo valor obtido na Tabela 1, não há correlação positiva forte entre a taxa de CVLI e o percentual da população vacinada no Ceará.

Tabela 1 – Correlação entre CVLI e vacinação

Variáveis investigadas	Coeficiente de Pearson
Taxa de CVLI por 100 mil habitantes	0,11246335478791318
Percentual de população vacinada	

5.1.2 Correlação Entre Taxa De Cvli e o Pib Per Capita Da População

A segunda questão de análise refere-se a entender o grau de associação entre as variáveis de taxa de CVLI por 100 mil habitantes e a renda per capita da população, visto que foi necessário entender se a produção de riqueza tem correlação com a incidência de crimes violentos letais e intencionais. Segundo o número obtido,

não é possível afirmar que há correlação entre as variáveis (veja Tabela 2).

Tabela 2 – Correlação entre CVLI e PIB

Variáveis investigadas	Coeficiente de Pearson
Taxa de CVLI por 100 mil habitantes	0,18813142622862755
PIB per capita do município	

5.1.3 Correlação Entre Taxa De População Vacinada e o Pib Per Capita Da População

Para analisar se há força de associação entre a taxa percentual de população e a produção de riqueza dentro de um município, calculado o seu coeficiente de correlação. De acordo com o coeficiente obtido e demonstrado na Tabela 3, não é possível identificar dependência entre as variáveis.

Tabela 3 – Correlação entre percentual de vacinação e PIB per capita

Variáveis investigadas	Coeficiente de Pearson
Percentual de população vacinada	0,17904913885122178
PIB per capita do município	

5.1.4 Correlação Entre O Número De Habitantes e a Taxa De Cvli

Com a finalidade de investigar se o número de habitantes está correlacionado com a taxa de CVLI nos municípios cearenses, foi calculado o coeficiente de Pearson entre as variáveis. Note que pelo valor obtido da Tabela 4 não existe correlação.

Tabela 4 – Correlação entre o número de habitantes e a taxa de CVLI

Variáveis investigadas	Coeficiente de Pearson
Número de habitantes	0,01865676809429713
Taxa de CVLI	

5.1.5 Correlação Entre o Índice De Desenvolvimento Humano e a Taxa De Cvli

Para aferir o nível de associação entre o índice de desenvolvimento humano de cada município e a taxa de CVLI, foi calculado o coeficiente de Pearson para a correlação. Contanto, conforme mostrado na Tabela 5 a correlação é fraca entre as variáveis.

Tabela 5 – Correlação entre o índice de desenvolvimento humano e a taxa de CVLI

Variáveis investigadas	Coeficiente de Pearson
Índice de desenvolvimento humano	0,2145879814729346
Taxa de CVLI	

5.2 Ranking Das Cidades Com Menores Percentuais De Vacinação e Maior Taxa De Cvli

Para traçar um paralelo entre os números obtidos de percentual de vacinação da população de cada município e a taxa de CVLI por 100 mil habitantes, foram construídos 2 *endpoints*, que retornam respectivamente os 10 maiores números de incidência de CVLI por 100 mil habitantes e menores percentuais de vacinação entre os municípios. Ao consumir os recursos e passar para uma tabela, pode-se verificar os seguintes dados, apresentados na Tabela 6:

Tabela 6 – Ranking das 10 cidades com maior taxa de CVLI por 10 mil habitantes

Posição	Município	Taxa de CVLI por 100 mil
1	São João do Jaguaribe	304,354
2	Monsenhor Tabosa	144,810
3	Chorozinho	138,026
4	Ibicuitinga	125,687
5	Forquilha	121,556
6	Quixeré	111,448
7	Guaiúba	98,084
8	Aquiraz	96,836
9	Itaitinga	95,704
10	Ibaretama	89,653

Como é possível notar a partir da visualização gerada nas Tabelas 6 e 7, somente uma cidade se repete nas duas listas, no caso Guaiúba, o que reforça a desassociação entre as variáveis de percentual de população vacinada e a taxa de CVLI por 100 mil habitantes.

Outras análises podem ser realizadas com os dados integrados e disponibilizados via API. O intuito de mostrar as correlações entre as variáveis e o ranqueamento foi oferecer ao leitor exemplos de como as análises e interpretação dos resultados

Tabela 7 – Ranking das 10 cidades com menores percentuais de vacinação

Posição	Município	Percentual de população vacinada
1	Nova Olinda	21,10
2	Guaiúba	27,57
3	Maranguape	28,25
4	Guaraciaba do Norte	28,83
5	Tarrafas	30,46
6	Aurora	30,85
7	Cariré	33,96
8	Graça	34,05
9	Morrinhos	34,91
10	Arneiroz	35,37

podem ser realizadas para os dados deste trabalho.

6 CONCLUSÕES E TRABALHOS FUTUROS

Foi delimitada a necessidade de demonstrar como dados públicos podem ser melhor disponibilizados, para fins de geração de conhecimento útil e melhor tomada de decisão das autoridades governamentais. Houve utilização de técnicas de extração e manipulação de dados, seguindo os processos de KDD. O presente trabalho extraiu, transformou, armazenou e disponibilizou os dados públicos em novas visualizações e meios de consumo com o intuito de gerar novo conhecimento útil.

As fontes de dados não dispunham de maneiras unificadas, homogêneas e filtráveis e que facilitassem a análise daqueles que viessem a ter interesse em consumi-las. É preciso que a disponibilização de dados públicos seja feita de maneira a facilitar o entendimento, manipulação e análise posterior por qualquer cidadão ou autoridade governamental, e não simplesmente seja um instrumento de cumprimento da lei.

Os objetivos previamente estabelecidos foram alcançados no desenvolvimento deste trabalho. Foi exposto como se pode disponibilizar os dados públicos de maneira mais otimizada, homogeneizando os dados, unificando e conectando-os. Foi dada a resposta para as questões problema, demonstrando também como ecossistemas secundários podem consumir o recurso através da API criada.

Para agregar ao proposto neste trabalho, se torna de forte importância em futuros trabalhos o acoplamento de funcionalidades que permitam a inserção de dados e sincronismo com fontes de dados públicos, trazendo visões dinâmicas, com dados atualizados, das origens agregadas. Além disso, torna-se interessante a demonstração de análises relativas a outros períodos, com outras fontes de dados e também com outras maneiras de correlacionar os dados, afim de investigar seus níveis de associação.

REFERÊNCIAS

ADRIAANS, P. **Data mining**. [S.l.]: Pearson Education India, 1996.

ASSIS, W. V. de; GOMIDE, J. V. B. Web scraping em dados públicos: método para extração de dados dos gastos públicos dos vereadores da câmara municipal de belo horizonte. **Informação & Informação**, v. 26, n. 4, p. 319–341, 2021.

BRASIL. **Lei de Acesso à Informação**. Brasília: [s.n.], 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>.

CERDEIRA, P. d. C.; MENDONÇA, M. M. d.; LAGOWSKA, U. G. **Políticas públicas orientadas por dados: os caminhos possíveis para governos locais**. [S.l.], 2020.

CHAMBERLIN, D. D.; BOYCE, R. F. Sequel: A structured english query language. In: **Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control**. [S.l.: s.n.], 1974. p. 249–264.

CORDEIRO, D. F. Avaliação da qualidade de dados públicos de extração e tratamento de produtos minerais. 2021. Disponível em: <https://bit.ly/3Ot7oAe>. Acesso em: 19 jun. 2022.

EAVES, D. **The Three Laws of Open Government**. 2009. Disponível em: <https://eaves.ca/2009/09/30/three-law-of-open-government-data/>. Acesso em: 19 jun. 2022.

FACEBOOK. **GraphQL**. 2016. Disponível em: <http://facebook.github.io/graphql/>. Acesso em: 19 jun. 2022.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996.

FERNANDEZ, G. P. *et al.* Orquestração de contêineres na nuvem: um modelo de segurança. Universidade Federal de Campina Grande, 2017.

FERREIRA, J.; MIRANDA, M.; ABELHA, A.; MACHADO, J. O processo etl em sistemas data warehouse. In: **INForum**. [S.l.: s.n.], 2010. p. 757–765.

FIELDING, R. T.; TAYLOR, R. N.; ERENKRANTZ, J. R.; GORLICK, M. M.; WHITEHEAD, J.; KHARE, R.; OREIZY, P. Reflections on the rest architectural style and"principled design of the modern web architecture"(impact paper award). In: **Proceedings of the 2017 11th joint meeting on foundations of software engineering**. [S.l.: s.n.], 2017. p. 4–14.

FORBES. **A importância dos dados abertos na luta contra a Covid-19 e no mundo pós-pandemia**. 2020. Disponível em: <https://forbes.com.br/forbes-tech/2020/05/a-importancia-dos-dados-abertos-na-luta-contra-a-covid-19-e-no-mundo-pos-pandemia/>. Acesso em: 05 jul. 2022.

HASURA. **What is Hasura?** 2018. Disponível em: <https://hasura.io/about/>. Acesso em: 19 jun. 2022.

JANNUZZI, P. d. M. **A importância da informação estatística para as políticas sociais no Brasil: breve reflexão sobre a experiência do passado para considerar no presente**. [S.l.]: SciELO Brasil, 2018.

KLEIN, R. H.; KLEIN, D. C. B.; LUCIANO, E. M. Identificação de mecanismos para a ampliação da transparência em portais de dados abertos: uma análise no contexto brasileiro. **Cadernos Ebape. br**, SciELO Brasil, v. 16, p. 692–715, 2018.

OPEN GOVERNMENT PARTNERSHIP. **About**. 2011. Disponível em: <https://www.opengovpartnership.org/about/>. Acesso em: 05 jul. 2022.

OPEN KNOWLEDGE FOUNDATION. **About**. 2021. Disponível em: <https://okfn.org/about/>. Acesso em: 19 jun. 2022.

RAUTENBERG, S.; BURDA, A. C.; SOUZA, L. de; DALL'AGNOL, J. M.; MICHELON, G.; HILD, T. A. Publicando dados na web de dados: Um relato de experiência na automatização dos processos de extração, transformação e carga de dados abertos provenientes do portal dados. gov. br. In: **XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XVIII ENANCIB)**. [S.l.: s.n.], 2017.

SILVA, F. H. R. Avaliação de desempenho de contêineres docker para aplicações do supremo tribunal federal. 2017.

W3C. **Manual dos Dados Abertos**. 2011. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 19 jun. 2022.