

# Inovação em saúde: a implementação de um *data lake* para armazenamento, sistematização e disponibilização de dados em saúde no Brasil

*Innovation in health: the implementation of a data lake for the storage, systematization and availability of health data in Brazil*

**Daniel do Prado Pagotto**

Doutorando em Administração pela Universidade de Brasília, UnB, Brasil.

ORCID: <https://orcid.org/0000-0001-6791-9511>

E-mail: [danielppagotto@ufg.br](mailto:danielppagotto@ufg.br)

**Wanderson da Silva Marques**

Graduado em Sistemas de Informação pelo Instituto Federal de Educação, Ciência e Tecnologia de Goiás, IFG, Brasil; Coordenador de Ciência de Dados na Secretaria de Estado da Saúde de Goiás, GO, Brasil.

ORCID: <https://orcid.org/0000-0003-2965-5759>

E-mail: [wdsmarques@gmail.com](mailto:wdsmarques@gmail.com)

**Denise Santos de Oliveira**

Doutora em Administração pela Universidade de Brasília, UnB, Brasil.

ORCID: <https://orcid.org/0000-0003-4981-119X>

E-mail: [deniseadm@hotmail.com](mailto:deniseadm@hotmail.com)

**Vicente da Rocha Soares Ferreira**

Doutor em Administração pela Universidade de Brasília; Professor associado da Faculdade de Administração, Ciências Contábeis e Ciências Econômicas da Universidade Federal de Goiás, Brasil.

ORCID: <https://orcid.org/0000-0002-1196-5778>

E-mail: [vicenterocha@ufg.br](mailto:vicenterocha@ufg.br)

**Vinicius Nunes de Azevedo**

Graduado em Administração de Empresas e Medicina pela Universidade Vila Velha, UVV, Brasil.

ORCID: <https://orcid.org/0000-0002-2112-5860>

E-mail: [viniciuscoach@gmail.com](mailto:viniciuscoach@gmail.com)

**Cândido Vieira Borges Júnior**

Doutor em Administração de Empresas pelo HEC Montreal, Canadá; Professor adjunto da Universidade Federal de Goiás, UFG, Brasil.

ORCID: <https://orcid.org/0000-0003-3362-4074>

E-mail: [candidoborges@ufg.br](mailto:candidoborges@ufg.br)

## Resumo

Este artigo tem como objetivo apresentar o problema relativo ao armazenamento, à sistematização e à disponibilização de dados em saúde no Brasil e uma solução inovadora: a implementação de um *data lake* com dados do setor de saúde. O *data lake* foi construído a partir de três etapas: (1) planejamento e priorização das bases de dados a serem importadas para o repositório; (2) extração, carregamento e tratamento dessas bases, com o apoio das ferramentas Apache Airflow e Dremio; e (3) aplicação do uso. Os resultados evidenciam a capacidade de a plataforma armazenar um grande volume de dados (*Big Data*), bem como de propiciar uma navegação intuitiva, facilitando a compreensão e o manuseio dos dados por analistas em saúde. Constata-se, ainda, que gestores públicos e pesquisadores reconhecem as contribuições da ferramenta para suas decisões e a potencialidade desta para o desenvolvimento de outras soluções de inteligência para a análise de dados da área de saúde. A solução apresentada visa contribuir para a gestão e o planejamento de políticas de saúde, permitindo o acesso de modo rápido e amplo a diversos dados que suportam a tomada de decisões na área de saúde com mais agilidade e segurança.

**Palavras-chave:** sistema de gestão de base de dados; análise de dados secundários; diretório de base de dados.

## Abstract

This article aims to present the problem related to the storage, systematization, and availability of health data in Brazil and an innovative solution: the implementation of a data lake with data from the health sector. The data lake was built from three steps: (1) planning and prioritizing the databases to be imported into the repository; (2) extraction, loading, and treatment of these bases, with the support of Apache Airflow and Dremio tools; and (3) application of use. The results show the platform's ability to store a large volume of data (Big Data), as well as to provide intuitive navigation, facilitating the understanding and the handling of data by health analysts. Note also that public managers and researchers recognize the tool's contributions to their decisions and its potential for the development of other intelligence solutions for data analysis in the health area. The solution presented aims to contribute to the management and planning of health policies, allowing quick and broad access to diverse data that support decision-making in the health area with more agility and security.

**Keywords:** database management system; secondary data analysis; database directory.

## 1. Introdução

A disponibilidade de dados confiáveis é crucial para que gestores de saúde possam tomar melhores decisões (Moutselos; Maglogiannis, 2020); assim, o tratamento, o gerenciamento e a análise adequada dos dados permitem a obtenção de informações fundamentais para a gestão dos serviços de saúde (Dash *et al.*, 2019). Em meio aos avanços tecnológicos, foram identificados aumentos expressivos no volume de dados registrados nos sistemas de informação em saúde (Shortreed *et al.*, 2019). Embora isso tenha representado muitas vantagens, tais como a vigilância em saúde, a análise da distribuição de profissionais entre os territórios e a prestação de cuidados de saúde, também gerou grandes desafios aos gestores, entre os quais a dificuldade de gerenciamento desse grande volume de dados (Kroezen; Van Hoegaerden; Batenburg, 2018; Moutselos; Maglogiannis, 2020).

Geralmente, esses dados são disponibilizados em repositórios isolados (Gamache; Kharrazi; Weiner, 2018). No Brasil, por exemplo, as bases do Cadastro Nacional de Estabelecimentos de Saúde (CNES) dispõem de dados sobre estabelecimentos de saúde, suas infraestruturas e os profissionais vinculados a eles. Ademais, o Sistema de Informação de Agravos de Notificação (Sinan), o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informações Hospitalares (SIH) fornecem dados de natureza epidemiológica. Por sua vez, as bases de projeções populacionais do Ministério da Saúde e do Instituto Brasileiro de Geografia e Estatística (IBGE) dispõem de dados sobre a demografia relativa ao sistema de saúde. Assim, observa-se que o problema a ser enfrentado pelo gestor, em uma análise macro, é encontrar, tratar, sistematizar, sintetizar e relacionar esse grande volume de dados advindo de diferentes fontes (Kroezen; Van Hoegaerden; Batenburg, 2018) e em diferentes formatos.

Para solucionar esse tipo de problema, é necessário reunir os dados em um único local. A integração dos dados de maneira estruturada permitiria aos gestores, planejadores e pesquisadores da saúde acessar, de modo rápido e fácil, um amplo conjunto de dados, bem como melhor visualizá-los, compreendê-los, e realizar análises a partir da associação entre elas (Dash *et al.*, 2019; Gamache; Kharrazi; Weiner, 2018). Nesse sentido, o objetivo deste artigo é apresentar o problema relativo ao armazenamento, à sistematização e à disponibilização de dados em saúde no Brasil e uma solução inovadora: a implementação de um *data lake* com dados públicos do setor de saúde.

O *data lake* é uma estrutura de armazenamento de dados que reúne informações de diversas fontes e aplica modelos analíticos para fornecer uma nova abordagem de interpretação, gerenciamento e análise aos usuários (Maini; Venkateswarlu; Gupta, 2018). A partir dele, gestores podem obter *insights* que permitam maior eficiência na gestão do sistema de saúde, em suas mais amplas ou específicas e complexas nuances.

A relevância deste estudo se sustenta na apresentação de todo o percurso de desenvolvimento do *data lake* e na descrição de todo o fluxo utilizado para a implementação de uma infraestrutura de dados estratégica e valiosa. Outra vantagem é a possibilidade desses mesmos procedimentos serem adotados em contextos diferentes. Para fins práticos, o resultado deste trabalho é um avanço, uma vez que objetiva consolidar, em uma fonte única e de fácil acesso, dados públicos, outrora pulverizados em múltiplas fontes e formatos que não possibilitavam a exploração completa de todo o seu potencial.

## 2. Sistemas brasileiros de informação e a consolidação dos dados

Nas últimas décadas, um conjunto de sistemas de informação foi implantado ou expandido no Brasil, o que ampliou a disponibilidade de informações para a gestão em saúde (Batista; Santana; Ferrite, 2019). Os gestores e os pesquisadores dispõem de uma rede de informações composta por dados demográficos, epidemiológicos, de monitoramento de programas de saúde, de quantidade de profissionais disponíveis, entre outros (Correia; Padilha; Vasconcelos, 2014).

Muitos sistemas de informação que geram dados em saúde de modo finalístico são administrados pelo Ministério da Saúde, tais como o e-SUS Notifica, o SIM, o Sistema de Informações Hospitalares do SUS (SIH/SUS), o Sistema de Informações sobre Nascidos Vivos

(Sinasc), o Sinan, o Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico (Vigitel), entre tantos outros. Tais sistemas armazenam dados que retratam as condições de saúde da população, fornecendo base a gestores e pesquisadores para o levantamento da demanda assistencial (Batista; Santana; Ferrite, 2019; Correia; Padilha; Vasconcelos, 2014). O Quadro 1 apresenta a finalidade de cada um desses sistemas.

Quadro 1 – Finalidades dos sistemas de saúde

<b>Bases de dados</b>	<b>Finalidades</b>
CNES	O Cadastro Nacional de Estabelecimentos de saúde é uma base que contempla os dados do estabelecimento de saúde, dos profissionais atuantes, da infraestrutura desses locais, entre outros aspectos.
e-SUS Notifica	Dispõe de notificações de síndrome gripal suspeita e notificação de covid-19.
SIM	Expõe dados sobre mortalidade.
SIA	Contempla os dados da prestação de serviços ambulatoriais.
SIH/SUS	Apresenta informações de internações hospitalares da rede própria ou conveniada ao Sistema Único de Saúde (SUS).
SINASC	Expõe informações sobre nascimentos.
SINAN	Apresenta dados de notificação compulsória de doenças e agravos.
VIGITEL	Apresenta dados acerca de doenças crônicas não transmissíveis, tais como diabetes, câncer e doenças cardiovasculares.
SIGTAP	Contempla dados associados a procedimentos realizados no SUS. Cada procedimento apresenta atributos complementares, como o valor e os profissionais aptos para realização.

Fonte: Batista, Santana e Ferrite (2019) e Brasil (2021).

Além desses sistemas, há outros que, apesar de não tratarem de saúde de modo finalístico ou direto, são importantes para o seu gerenciamento. O IBGE, por exemplo, apresenta projeções populacionais e características dos municípios que podem auxiliar no processo de gestão. Por sua vez, o Censo da Educação Superior, realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), permite identificar a formação de recursos humanos em saúde (Machado; Ximenes Neto, 2018).

Assim, as iniciativas de sistematização de dados são diversas, mas esses dados são apresentados de modo fragmentado e, muitas vezes, são de difícil compreensão e acesso para o

gestor (Correia; Padilha; Vasconcelos, 2014). Sabe-se, contudo, que a forma como os dados são apresentados influencia a compreensão e a formação de *insights* (Fernandes *et al.*, 2020). Nesse sentido, o uso de ferramentas de monitoramento das informações, tais como *dashboards*, torna-se relevante. Os *dashboards* são painéis que exibem informações diversas a partir de gráficos claros, objetivos, completos e passíveis de customizações por meio da utilização de filtros, que podem ser incorporados, permitindo a visualização de um conjunto de dados de forma simples, o que auxilia na tomada de decisões (Knafllic, 2015). Promovem, ainda, a apresentação de tendências e a associação entre diferentes variáveis, simplificando a análise e potencializando a construção ou reconstrução de estratégias de modo preciso, customizado e ágil (Nijkamp; Kourtit, 2022).

Embora os *dashboards* auxiliem na visualização de dados, a sua apresentação de modo fragmentado, isto é, a partir de diferentes sítios, pode produzir dificuldades na sua devida identificação pelo gestor, bem como desfavorecer a devida análise das informações. Dessa forma, o agrupamento de informações de fontes diversas em um único ambiente pode auxiliar os gestores e os pesquisadores de saúde na sua identificação e na obtenção de uma visão mais ampla dos dados (Dash *et al.*, 2019; Gamache; Kharrazi; Weiner, 2018).

Algumas iniciativas para a consolidação de dados em sítio único são evidenciadas no Brasil, tais como os esforços da organização não governamental Base dos Dados, a Rede Nacional de Dados em Saúde (RNDS) e a Plataforma de Análise de Dados do Governo Federal (GovData). A Base de Dados foi iniciada em 2019 e visa facilitar o acesso a dados públicos, de modo geral, por meio de bibliotecas das linguagens de programação *R* e *Python* ou através da ferramenta *BigQuery* (Base dos Dados, 2022).

A RNDS é um programa instituído pelo Ministério da Saúde, desde maio de 2020, para integrar dados de diversos atores de todo o país em uma plataforma nacional única (interoperabilidade de sistemas de saúde), incluindo dados de laboratórios, centros de pesquisa e desenvolvimento, farmácias, profissionais de saúde, atendimento de urgência e emergência, entre outros, permitindo o compartilhamento de informações da assistência à saúde nos setores público e privado (Coutinho; Neves; Lopes, 2021; Brasil, 2020).

A GovData, por sua vez, é uma iniciativa do Serviço Federal de Processamento de Dados (Serpro) e da Empresa de Tecnologia e Informações da Previdência Social (Dataprev) que visa dispor de um ambiente unificado para análise, compartilhamento e cruzamento de dados governamentais. A plataforma utiliza diversas bases de dados do governo federal, como

o Cadastro de Pessoa Física (CPF), o Sistema Integrado de Administração de Recursos Humanos (Siape), o Sistema Integrado de Administração Financeira (Siafi) e o Registro Nacional de Veículos Automotores (Renavam). Esse *data lake* está disponível para o acesso de empresas (mercado privado) e órgãos públicos (em todas as esferas) (Araújo; Zullo; Torres, 2020; Brasil, 2023).

Tais esforços da organização não governamental Base dos Dados, a RNDS e o GovData contribuem para a sistematização dos dados em saúde. O presente estudo avança em relação a estes ao incluir a *Analytics Layer*, que permite a construção conjunta de consultas por todos os usuários do *data lake*, proporcionando transparência aos códigos e o reaproveitamento dos scripts por outros que necessitem de análises semelhantes.

A implementação de um novo *data lake* da área da saúde pode contribuir para suprir as limitações identificadas, visto que ele armazena, em tempo real, dados de fontes diversas em seu formato bruto. Cada fonte está associada a um identificador único, e a construção de um *data lake* demanda um repositório de metadados capaz de registrar um alto nível de informações sobre entidades de dados. Nele, é possível apresentar dados a partir de diversas ferramentas de análise e visualização de dados, o que auxilia a compreensão. Além do registro de dados, é possível aplicar modelos analíticos para associar dados de uma mesma base ou de bases diferentes (Maini; Venkateswarlu; Gupta, 2018).

### 3. Procedimentos metodológicos

A construção da solução ocorreu ao longo do ano de 2021 e contou com três etapas, conforme ilustrado na Figura 1.

Figura 1 – Etapas para construção do *data lake*.

Fonte: Elaborado pelos autores (2023)

A primeira etapa foi destinada ao planejamento do projeto e à priorização das bases de dados públicas a serem importadas para o *data lake*. Esta última atividade foi realizada por meio da consulta a três pesquisadores – um mestre e dois doutores – com experiência em análise de dados em saúde e responsáveis por liderar três projetos diferentes na área. Os especialistas foram acessados no intuito de listar bases de dados públicas úteis para os projetos nos quais estavam envolvidos. Como resultado, foi formada uma lista de 27 bases oriundas de diferentes fontes, como Datasus, IBGE, Ministério da Educação (MEC), Ministério do Trabalho e Emprego (MTE), Agência Nacional de Saúde Suplementar (ANS), entre outras. Identificada a relação de bases, os pesquisadores indicaram quais seriam aplicadas nos respectivos projetos, o que permitiu estabelecer uma ordem de priorização para a inclusão no *data lake*. Seriam priorizadas na etapa 2 aquelas que seriam utilizadas nas três iniciativas, por exemplo.

A segunda etapa correspondeu à condução de ciclos de extração, tratamento e carregamento das bases priorizadas. Portanto, para cada base, os dados foram extraídos de seus sítios originais e tratados de modo a padronizarem-se os diferentes formatos de dados em arquivos do tipo *parquet*<sup>1</sup>, formato este com maior grau de compressão de volume se comparado a outros tipos, tais como *comma separated values* (csv). Além disso, para as bases com atualização periódica, foram construídos scripts nas linguagens R e *Python* para automatizar

<sup>1</sup> *Parquet* é um formato de arquivo que armazena grandes volumes de dados de modo colunar e compactado. Em comparação com outros formatos, ele demonstra melhor desempenho para compactação, reduzindo o tamanho dos dados no disco (Vohra, 2016).

a extração, transformação e carregamento dos dados (em inglês, ETL). A implementação de rotinas de ETL pode ser realizada de diversas formas, desde códigos utilizando linguagens de programação e SQL (*Structured Query Language*) até ferramentas *no/low-code*<sup>2</sup>, como *Pentaho Data Integration*, *Talend* e *Oracle Data Integrator*.

Optou-se pela implementação das rotinas ETL utilizando *Python* e *R*, pois ambas são linguagens de código aberto, têm grande suporte da comunidade, apresentam desempenho satisfatório e são amplamente utilizadas para processar dados massivos. O uso de ambas linguagens também facilitou o processo de controle de versões. O processo de orquestração, ou seja, de controle de *logs*, agendamentos e gerenciamento das execuções, foi realizado por meio da plataforma de orquestração de fluxo de dados *Apache Airflow*<sup>3</sup>.

A Figura 2 ilustra a arquitetura do *data lake*, com as tecnologias que mantêm o funcionamento da ferramenta. Os dados foram acessados a partir das múltiplas fontes, sob diferentes formatos e carregados em estado bruto (*raw data*) em uma primeira camada de dados. Atualmente, têm-se 43 bases, sendo elas da área de saúde (como do CNES, e-SUS, SIM e outras), do IBGE, do MEC e de outras fontes. Uma segunda camada – *analytics layer* – foi desenvolvida para permitir que o usuário realize tratamento e consultas de dados e os salve, o que garante a transparência na construção de consultas e seu reaproveitamento. Por fim, ainda há duas camadas: uma de catálogo, que descreve os dados presentes no *data lake*; outra de segurança (autenticação, controle de usuários, registro de *logs*, etc.).

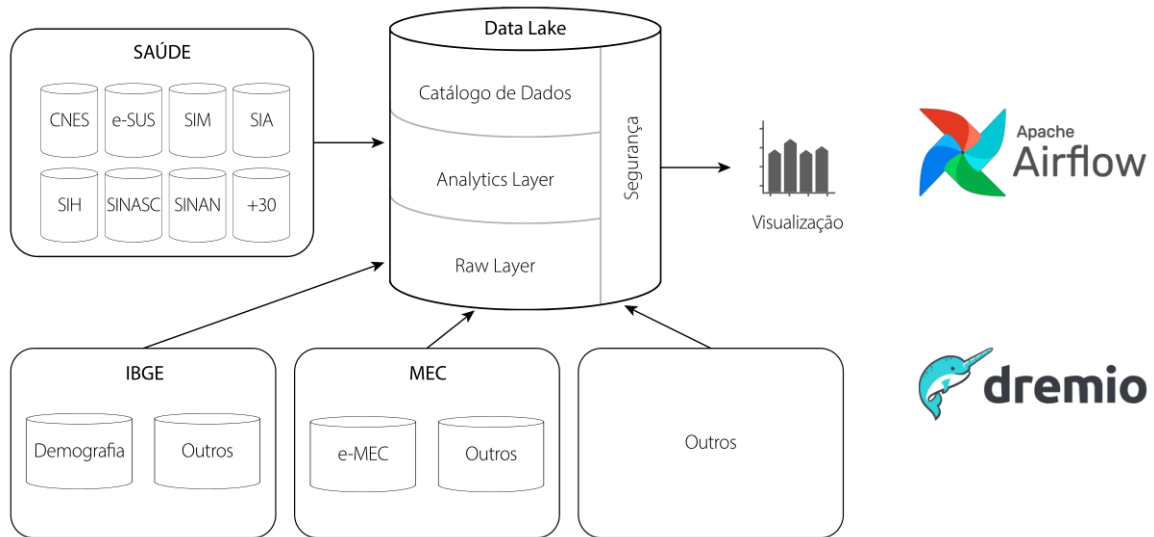
---

<sup>2</sup> As ferramentas *no-code* permitem a criação de aplicativos sem a necessidade de programação e as ferramentas *low-code*, com menos codificação manual, assim, tornam o processo mais acessível a pessoas com menos experiência em programação (Rokis; Kirikova, 2022).

<sup>3</sup> *Apache Airflow* é uma ferramenta de código aberto para automatizar fluxos de trabalhos. Ela monitora a execução de tarefas em pipelines de dados e processos (Harenslak; Rwitter, 2021).



Figura 2 – Estrutura de funcionamento do *data lake*.



Fonte: Elaborado pelos autores (2023)

A última etapa consistiu na disponibilização da ferramenta para os usuários dos três projetos mencionados, para que pudessem testá-la e relatar possíveis erros, para, em sequência, consumi-la. Nessa etapa, também foi realizado um treinamento, que contou com a participação de dez pesquisadores com experiência em análise de dados e em projetos de saúde, no intuito de apresentar o funcionamento da ferramenta, as formas de acesso via linguagens SQL, R e *Python* e a interface do *Dremio*.

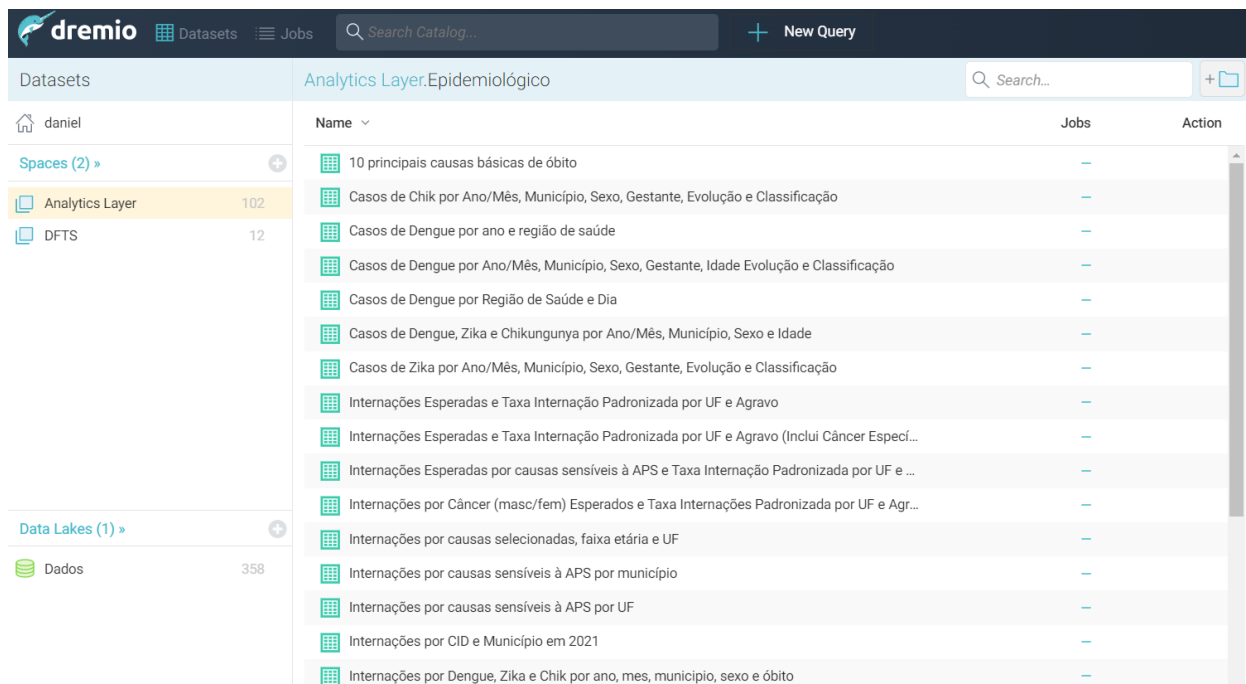
Uma vez difundido o uso do *data lake*, essa estrutura de armazenamento de dados passou a ser utilizada como referência de acesso a banco de dados, permitindo um uso mais eficiente tanto para cálculos rotineiros quanto para a construção de plataformas de *Business Intelligence*, uma vez que o *Dremio* tem a funcionalidade de integração direta a softwares dessa natureza, como o *Microsoft Powerbi* e o *Tableau*.

#### 4. Resultados

Ao final do primeiro ano do projeto, o *data lake* já contava com mais de 300 gigabytes de dados e 43 bases, subdivididas em 415 tabelas; ao todo, cinco projetos e 28 pesquisadores que atuam em iniciativas sobre gestão do trabalho e educação na saúde consomem dados diretamente dele.

O *data lake* trouxe vantagens, como a padronização na forma de acessar dados e a eficiência decorrente de tratamentos prévios realizados por meio de linguagem SQL, utilizando a interface do *Dremio* (Figura 3) e linguagens de programação.

Figura 3 – Interface de acesso a dados via *Dremio* com SQL.



Name	Jobs	Action
10 principais causas básicas de óbito	–	
Casos de Chik por Ano/Mês, Município, Sexo, Gestante, Evolução e Classificação	–	
Casos de Dengue por ano e região de saúde	–	
Casos de Dengue por Ano/Mês, Município, Sexo, Gestante, Idade Evolução e Classificação	–	
Casos de Dengue por Região de Saúde e Dia	–	
Casos de Dengue, Zika e Chikungunya por Ano/Mês, Município, Sexo e Idade	–	
Casos de Zika por Ano/Mês, Município, Sexo, Gestante, Evolução e Classificação	–	
Internações Esperadas e Taxa Internação Padronizada por UF e Agravo	–	
Internações Esperadas e Taxa Internação Padronizada por UF e Agravo (Inclui Câncer Especí...	–	
Internações Esperadas por causas sensíveis à APS e Taxa Internação Padronizada por UF e ...	–	
Internações por Câncer (masc/fem) Esperados e Taxa Internações Padronizada por UF e Agr...	–	
Internações por causas selecionadas, faixa etária e UF	–	
Internações por causas sensíveis à APS por município	–	
Internações por causas sensíveis à APS por UF	–	
Internações por CID e Município em 2021	–	
Internações por Dengue, Zika e Chik por ano, mes, municipio, sexo e óbito	–	

Fonte: Elaborado pelos autores (2023).

O depoimento abaixo registra o comentário de um dos usuários do *data lake*, Usuário 1 (2023):

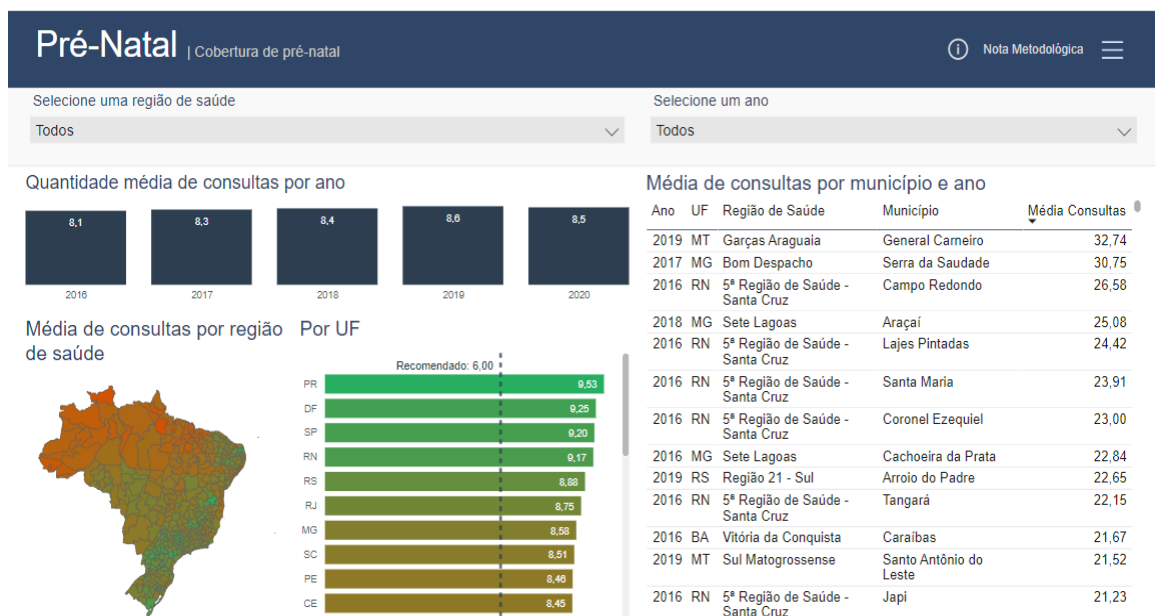
O *data lake* trouxe muitos benefícios para quem trabalha e pesquisa sobre os temas relacionados às suas bases de dados. Primeiro, é uma fonte única de dados. Existem bases que podem ser acessadas via protocolo de transferência de arquivo (ftp, do inglês *file transfer protocol*) do Datasus, por meio da ferramenta para tabulação de dados Tabnet ou por meio de um site do Ministério da Saúde. Por alguma questão de tratamento, estas fontes podem apresentar alguns números – ainda que pequenos – diferentes. Portanto, uniformizar a fonte de acesso foi o primeiro ganho. Além disso, ele padronizou todos os arquivos em um formato único, que pode ser acessado via comandos SQL. Assim, o pesquisador não precisa baixar diversos arquivos em diferentes formatos, como csv ou planilhas eletrônicas. Basta acessar via SQL e extrair aquelas variáveis/atributos que necessitará para suas análises. Decorrente desta última

vantagem está o uso do SQL, que permite um tratamento prévio dos dados de um modo mais eficiente do que por meio do acesso às bases brutas e o tratamento via bibliotecas *tidyverse* ou *pandas* das linguagens R e *Python*, respectivamente. A última vantagem está na função da *analytics layers*, que permite reaproveitar as consultas já construídas.

Os resultados do projeto foram apresentados em duas ocasiões diferentes para gestores de nível estratégico<sup>4</sup>, o reitor da universidade à qual o projeto está vinculado e um gestor de nível tático<sup>5</sup> da administração pública federal. Diante da potencialidade e importância reforçada nesses encontros, foi proposta a continuidade da iniciativa e a incorporação de outras bases de dados, difundindo-se informações no formato de *dashboards*.

Diante do exposto e considerando que um dos propósitos do *data lake* foi centralizar dados a fim de aplicá-los em sistemas e painéis de visualização de dados, a equipe do projeto avançou na criação de *dashboards* sobre dados em saúde. A Figura 4 ilustra um deles.

Figura 4 – *Dashboard* sobre cobertura de pré-natal.



Fonte: Elaborado pelos autores (2023).

<sup>4</sup> É o mais alto nível de decisão dentro de uma organização. Ele tem como foco as decisões de longo prazo, que afetam a organização como um todo (Sobral; Peci, 2013).

<sup>5</sup> É o nível intermediário de decisões. Ele tem como foco implementar as estratégias do alto nível, estabelecendo planos e ações de médio prazo (Sobral; Peci, 2013).

## 5. Discussão

O uso de dados secundários oriundos de diferentes bases de dados é amplamente utilizado na área de saúde, seja na formulação de políticas de saúde, na gestão dos serviços ou no desenvolvimento de pesquisas (Ferreira *et al.*, 2020). No entanto, por vezes, tais dados são apresentados de modo fragmentado (Coelho Neto; Chioro, 2021; Correia; Padilha; Vasconcelos, 2014). Superar a fragmentação de dados na área de saúde é um desafio não só para o Brasil (Pinto; Freitas; Figueiredo, 2018).

Uma das formas mais utilizadas de acesso aos dados da área de saúde é por meio do Tabnet/Tabwin (sistemas de informações do Datasus que disponibilizam dados do SUS. No entanto, por mais que tenha sido um avanço para a publicidade de dados em saúde, é uma ferramenta desenvolvida há cerca de 30 anos (Coelho Neto; Chioro, 2021) e que apresenta limitações importantes, incluindo a impossibilidade de aplicação de agregações não compatíveis às funcionalidades da plataforma. Nesse sentido, algumas experiências são importantes para a disseminação de microdados (dados brutos, com menor granularidade), como o desenvolvimento do pacote da linguagem R microdatasus (Saldanha; Bastos; Barcellos, 2019).

Como mais uma estrutura de armazenamento de dados, o *data lake* apresentado e descrito neste estudo permite que inúmeros dados sejam acessados a partir de uma única fonte, viabilizando, assim, o aumento da confiabilidade e usabilidade das plataformas de acesso aos dados. Além disso, a inclusão de camada analítica com a consolidação de consultas e análises constitui-se um passo fundamental para garantir maior transparência por meio de um paradigma de dados e materiais abertos (Miguel *et al.*, 2014), além de contribuir para a produtividade através do reaproveitamento de consultas construídas por outros pesquisadores.

Diante da ampla variedade de dados disponíveis, é importante que se busque ferramentas que sintetizem e sistematizem tal volume, garantindo-lhes acessibilidades e navegações intuitivas, condições fundamentais para a efetiva garantia de transparência proposta com a disponibilização do acesso aos dados. Nesse sentido, a visualização de dados por meio de painéis interativos pode ser uma estratégia que facilita a absorção de informações (Dash *et al.*, 2018), contribuindo, em última instância, para uma melhor tomada de decisão pelos gestores (Ifitikhar *et al.*, 2019), o que garante eficiência, eficácia e efetividade ao trabalho de gestores e pesquisadores.

## 6. Conclusão

Há avanços constantes sobre a disponibilidade de dados públicos decorrentes de sistemas governamentais. Apesar desses avanços decorrentes de esforços públicos e do terceiro setor, iniciativas de difusão de dados e informações geradas a partir deles devem ser incentivadas. Nesse sentido, este artigo apresenta o problema relativo ao armazenamento, à sistematização e à disponibilização de dados em saúde no Brasil e uma solução inovadora, a implementação de um *data lake* com dados do setor de saúde que contribui para a difusão de dados, a transparência de análises, o apoio gerencial e apoio à pesquisa.

Além dos benefícios inerentes ao produto listados acima, o artigo descreve todo o processo de construção da plataforma, apresentando ferramentas e processos sistematizados adotados, o que permite o desenvolvimento de estruturas semelhantes para outras finalidades.

A pesquisa limita-se pelo uso exclusivo de dados secundários públicos, os quais são amplamente utilizados em pesquisas e na tomada de decisão. Todavia, isso não exime os gestores de sistemas de informação e bases de dados de buscarem melhorias constantes nesse universo. Além disso, novos métodos que assegurem uma constante melhoria na qualidade dos dados podem ser incorporados, como aprimoramento do preenchimento na fonte, tratamento e uso de instrumentos de coleta que assegurem a captura de dados relevantes.

Ainda, o acesso a microdados, de modo geral, consiste em uma limitação deste estudo. Nesse sentido, mais esforços devem ser empregados para a disponibilização de microdados de certas bases, como dos componentes do e-SUS AB. Tais questões vão além do escopo dessa ferramenta, mas, por tangenciar o principal insumo deste trabalho – os dados – não podem deixar de ser alertadas. Espera-se que essas medidas, em conjunto com a maior abertura no acesso a dados, possam ampliar as pesquisas no campo da saúde, bem como aprimorar o gerenciamento de organizações e serviços de saúde, além de conceder um grande apoio ao trabalho de pesquisadores da área de saúde pública.

## Referências

ARAÚJO, V. S.; ZULLO, B. A.; TORRES, M. Big data, algoritmos e inteligência artificial na Administração Pública: reflexões para a sua utilização em um ambiente democrático. **A&C-Revista de Direito Administrativo & Constitucional**, Curitiba, v. 20, n. 80, p. 241-261, 2020. Disponível em: <http://dx.doi.org/10.21056/aec.v20i80.1219>. Acesso em: 10 out. 2023.

BASE DOS DADOS. **Quem somos**. 2022. Disponível em: <https://basedosdados.org/quem-somos>. Acesso em: 09 abr. 2022.

BATISTA, A. G.; SANTANA, V. S.; FERRITE, S. Registro de dados sobre acidentes de trabalho fatais em sistemas de informação no Brasil. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 24, p. 693-704, 2019. Disponível em: <https://www.scielo.org/article/csc/2019.v24n3/693-704/pt/>. Acesso em: 09 abr. 2022.

BRASIL. **Contratar plataforma de análise de dados para suporte a políticas públicas** (GovData). 05 jan. 2023. Disponível em: <https://www.gov.br/pt-br/servicos/contratar-plataforma-de-analise-de-dados-para-suporte-a-politicas-publicas-govdata>. Acesso em: 10 de out. 2023.

BRASIL. MINISTÉRIO DA SAÚDE. Gabinete do Ministro. **Portaria nº 1.434, de 28 de maio de 2020**. Disponível em: <https://www.in.gov.br/en/web/dou/-/portaria-n-1.434-de-28-de-maio-de-2020-259143327>. Acesso em: 07 de abr. 2022.

BRASIL. MINISTÉRIO DA SAÚDE. **Sistemas de informação em saúde**. 2021. Disponível em: <https://www.gov.br/saude/pt-br/composicao/svs/vigilancia-de-doencas-cronicas-nao-transmissiveis/sistemas-de-informacao-em-saude>. Acesso em: 07 de abr. 2022.

COELHO NETO, G. C.; CHIORO, Arthur. Afinal, quantos sistemas de informação em saúde de base nacional existem no Brasil? **Cadernos de Saúde Pública**, Rio de Janeiro, v. 37, n. 7, jul. 2021, e00182119. Disponível em: <https://www.scielo.org/article/csp/2021.v37n7/e00182119/>. Acesso em: 12 abr. 2022.

CORREIA, L. O. D. S.; PADILHA, B. M.; VASCONCELOS, S. M. L. Métodos para avaliar a completude dos dados dos sistemas de informação em saúde do Brasil: uma revisão sistemática. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 19, p. 4467-4478, 2014. Disponível em: <https://www.scielo.br/j/csc/a/HGyrfBHWLXMd3mz74HCcvpy/abstract/?lang=pt>. Acesso em: 14 abr. 2022.

COUTINHO, L. R.; NEVES, H. P. O. D. E.; LOPES, L. C. Abordagens sobre computação na nuvem: uma breve revisão sobre segurança e privacidade aplicada a e-saúde no contexto do Programa Conecte SUS e Rede Nacional de Dados em Saúde (RNDS). **Brazilian Journal of Development**, Curitiba, v. 7, n. 4, p. 35152-35170, abr. 2021. Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/view/27732>. Acesso em: 14 abr. 2022.

DASH, S. *et al.* Big data in healthcare: management, analysis and future prospects. **Journal of Big Data**, v. 6, n. 1, p. 1-25, jun. 2019. Disponível em: <https://link.springer.com/article/10.1186/s40537-019-0217-0>. Acesso em: 14 abr. 2022.

FERNANDES, A. M. R. *et al.* A relevância dos dashboards para a gestão da saúde na pandemia causada pelo COVID-19. **Brazilian Journal of Development**, Curitiba, v. 6, n. 6, p. 39263-39274, jun. 2020. Disponível em:

<https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/11931>. Acesso em: 09 abr. 2022.

FERREIRA, J. E. D. S. M. *et al.* Sistemas de informação em saúde no apoio à gestão da atenção primária à saúde: revisão integrativa. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, Rio de Janeiro, v. 14, n. 4, p. 970-982, out./dez. 2020. Disponível em: <https://www.arca.fiocruz.br/handle/icict/45028>. Acesso em: 14 abr. 2022.

GAMACHE, R.; KHARRAZI, H.; WEINER, J. P. Public and population health informatics: the bridging of big data to benefit communities. **Yearbook of medical informatics**, v. 27, n. 1, p. 199-206, 2018. Disponível em: <https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0038-1667081>. Acesso em: 09 abr. 2022.

HARENSLAK, B. P.; RUITER, J. **Data pipelines with Apache airflow**. New York: Simon and Schuster, 2021.

IFTIKHAR, A. *et al.* Role of dashboards in improving decision making in healthcare: Review of the literature. In: EUROPEAN CONFERENCE ON COGNITIVE ERGONOMICS – ECCE'19, 31. 2019, Belfast. **Proceedings...** Belfast: ACM, 2019. p. 215-219. Disponível em: <https://dl.acm.org/doi/abs/10.1145/3335082.3335109>. Acesso em: 12 abr. 2022.

KNAFLIC, C. N. **Storytelling with data: a data visualization guide for business professionals**. New Jersey: John Wiley & Sons, 2015. Disponível em: [https://books.google.com.br/books?hl=pt-BR&lr=&id=retRCgAAQBAJ&oi=fnd&pg=PR9&dq=A+data+visualization+guide+for+business+professionals&ots=KpeLBnMy7\\_&sig=I9Ab4ITpts7IaZFvSUqJA4OIgHE#v=onepage&q=A%20data%20visualization%20guide%20for%20business%20professionals&f=false](https://books.google.com.br/books?hl=pt-BR&lr=&id=retRCgAAQBAJ&oi=fnd&pg=PR9&dq=A+data+visualization+guide+for+business+professionals&ots=KpeLBnMy7_&sig=I9Ab4ITpts7IaZFvSUqJA4OIgHE#v=onepage&q=A%20data%20visualization%20guide%20for%20business%20professionals&f=false). Acesso em: 12 abr. 2022.

KROEZEN, M.; VAN HOEGAERDEN, M.; BATENBURG, R. The joint action on health workforce planning and forecasting: results of a european programme to improve health workforce policies. **Health Policy**, v. 122, n. 2, p. 87-93, fev. 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016885101730341X>. Acesso em: 12 abr. 2022.

MACHADO, M. H.; XIMENES NETO, F. R. G. Gestão da educação e do trabalho em saúde no SUS: trinta anos de avanços e desafios. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 23, n.6, p. 1971-1979, jun. 2018. Disponível em: <https://www.scielosp.org/article/csc/2018.v23n6/1971-1979/>. Acesso em: 12 abr. 2022.

MAINI, E.; VENKATESWARLU, B.; GUPTA, A. Data lake-an optimum solution for storage and analytics of Big Data in cardiovascular disease prediction system. **International Journal of Computational Engineering & Management**, v. 21, n. 6, p. 33-39, 2018. Disponível em: [http://ijcem.org/papers112018/ijcem\\_112018\\_05.pdf](http://ijcem.org/papers112018/ijcem_112018_05.pdf). Acesso em: 20 abr. 2022.

MIGUEL, E. *et al.* Promoting transparency in social science research. **Science**, v. 343, n. 6166, p. 30-31, jan. 2014. Disponível em: <https://www.science.org/doi/full/10.1126/science.1245317>. Acesso em: 20 abr. 2022.

MOUTSELOS, K.; MAGLOGIANNIS, I. Evidence-based public health policy models development and evaluation using big data analytics and web technologies. **Medical Archives**, v. 74, n. 1, p. 47-53, fev. 2020. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164729/>. Acesso em: 24 abr. 2022.

NIJKAMP, P.; KOURTIT, K. Place-specific corona dashboards for health policy: design and application of a ‘Dutchboard’. **Sustainability**, v. 14, n. 2, p. 836, jan. 2022. Disponível em: <https://www.mdpi.com/2071-1050/14/2/836>. Acesso em: 09 abr. 2022.

PINTO, L. F.; FREITAS, M. P. S. D.; FIGUEIREDO, A. W. S. A. D. Sistemas nacionais de informação e levantamentos populacionais: algumas contribuições do Ministério da Saúde e do IBGE para a análise das capitais brasileiras nos últimos 30 anos. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 23, p. 1859-1870, jun. 2018. Disponível em: <https://www.scielo.org/article/csc/2018.v23n6/1859-1870/pt/>. Acesso em: 25 abr. 2022.

ROKIS, K.; KIRIKOVA, M. Challenges of low-code/no-code software development: a literature review. *In: INTERNATIONAL CONFERENCE ON BUSINESS INFORMATICS RESEARCH*, 21., 2022, Rostock. **Perspectives in business informatics research**. Cham: Springer International, 2022. p. 3-17.

SALDANHA, R. D. F.; BASTOS, R. R.; BARCELLOS, C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cadernos de Saúde Pública**, Rio de Janeiro, v. 35, n. 9, set. 2019. Disponível em: <https://www.scielo.br/j/csp/a/gdJXqcrW5PPDHX8rwPDYL7F/>. Acesso em: 24 abr. 2022.

SHORTREED, S. M. *et al.* Challenges and opportunities for using big health care data to advance medical science and public health. **American Journal of Epidemiology**, v. 188, n. 5, p. 851-861, maio 2019. Disponível em: <https://academic.oup.com/aje/article/188/5/851/5381891?login=true>. Acesso em: 24 abr. 2022.

SOBRAL, F.; PECI, A. **Administração: teoria e prática no contexto brasileiro**. 2. ed. São Paulo: Pearson, 2013.

VOHRA, D. **Practical hadoop ecosystem: a definitive guide to hadoop-related frameworks and tools**, Apache parquet, p. 325-335, set. 2016. Disponível em: [https://link.springer.com/chapter/10.1007/978-1-4842-2199-0\\_8](https://link.springer.com/chapter/10.1007/978-1-4842-2199-0_8). Acesso em: 10 out. 2023.

Artigo submetido em: 19 jun. 2023

Artigo aceito em: 05 fev. 2024