UNIVERSIDADE FEDERAL DA PARAÍBA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Alberis Ricardo Tavares Filho

Uma plataforma de Business Intelligence para análise de dados da COVID-19 no Brasil

Rio Tinto
2021

Alberis Ricardo Tavares Filho

Uma plataforma de Business Intelligence para análise de dados da Covid-19 no Brasil

Trabalho de conclusão de curso para aprovação no curso de Bacharelado em Sistemas de Informação da Universidade Federal da Paraíba.

Orientador: Prof. Dr. Marcus Williams Aquino de Carvalho

Rio Tinto
2021

Abstract: *This article follows the procedures of the development of a Business Intelligence platform for the analysis of COVID-19 data, from the beginning of the pandemic in Brazil to the present day, through the theoretical foundation, detailing of the ETL process, of the modeling a dimensional database(STAR SCHEMA), integrating the database with a business intelligence tool, and displaying the key features of the platform. The project used databases made available by the federal government, referring to COVID-19..*

Resumo: *Este artigo tem por finalidade detalhar os procedimentos do desenvolvimento de uma plataforma de Business Intelligence para análise dos dados da COVID-19, desde o início da pandemia no Brasil até os dias atuais, através da fundamentação teórica, detalhamento do processo de ETL, da modelagem de um banco de dados dimensional(STAR SCHEMA), da integração do banco de dados com uma ferramenta de business intelligence e da exibição das funcionalidades principais da plataforma. O projeto utilizou as bases de dados disponibilizadas pelo governo federal, referentes à COVID-19.*

1. Introdução

Em 26 de fevereiro de 2020, foi registrado o primeiro caso de COVID-19 no Brasil[1], causando muitas dúvidas sobre o cenário que se sucederia a partir dali e pânico entre a população. Então, como em todo o mundo, o Brasil começou a contabilizar os dados referentes a casos confirmados, casos suspeitos, óbitos, locais onde eram notificados esses casos, entre outros indicadores sobre o avanço do vírus no país. Até novembro de 2021 já foram contabilizados mais de 614 mil óbitos em função da doença no Brasil. Porém, segundo Mellis (2021), o número de óbitos diários vem caindo à medida que o número de brasileiros vacinados aumenta. A média de óbitos diários que ultrapassou 3.000 em abril de 2021, no pico da pandemia no país, reduziu para menos de 300 em novembro de 2021 após avanço na vacinação.

O governo federal disponibiliza dados referentes à pandemia da COVID-19 no Brasil, como os dados da base de Notificações de Síndrome Gripal (SG) e Síndrome Respiratória Aguda Grave (SRAG), porém esses dados não estão estruturados de maneira que possibilite uma análise detalhada das informações sem que o usuário tenha um conhecimento prévio em programação, para desenvolver uma solução que transforme os dados das tabelas do governo em informação. Além disso, os dados são disponibilizados através de bases de dados distintas e heterogêneas. Dessa forma, é muito difícil para pesquisadores e outros interessados, como cidadãos, realizar uma análise desses dados, cruzando informações de diferentes fontes, filtrando, agregando, extraindo e visualizando as informações de acordo com a necessidade de detalhamento da informação.

[1]https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2020/02/brasil-confirma-primeiro-caso-do-novo-coronavirus

Tendo em vista as dificuldades de se extrair informações relevantes dos dados disponibilizados pelo governo, esse projeto tem como objetivo desenvolver um sistema que disponibilize, para o pesquisadores e outros agentes interessados, uma plataforma de Business Inteligence (BI) que cruze os dados da COVID-19 de várias fontes e que possibilite aos stakeholders agregar, extrair e visualizar os dados, facilitando e otimizando sua análise.

2. Arquitetura do sistema

Para atender aos objetivos, foi desenvolvido um sistema composto por um processo de ETL (Extract, Transform and Load), que é um processo que extrai informações de bases de dados, as processa, transforma e, por fim, carrega essas informações em outra base de dados chamada *Data Warehouse* (DW), que consiste em um repositório único de dados consolidados e organizados em uma ferramenta de Business Intelligence (BI) (Primak, 2008) (Ferreira et al. 2010), como mostra a Figura 1.

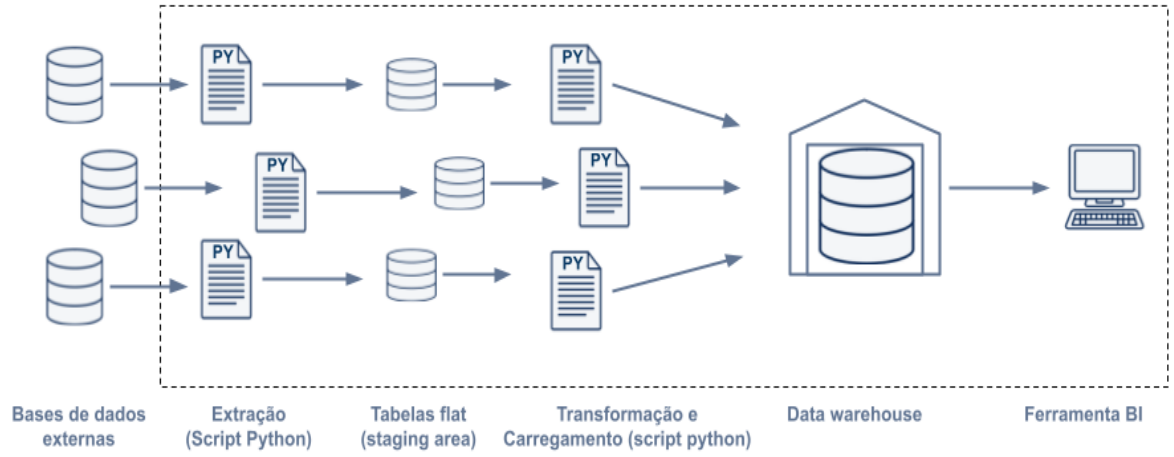


Figura 1. Ilustração do fluxo dos dados no sistema.

2.1. Processo de ETL

O processo de ETL é dividido em 3 partes: extração, que coleta as informações em bases de dados externas; transformação, que aplica regras de negócios, limpa os dados e faz a estruturação deles em um modelo dimensional; e, por último, o carregamento dos dados, que consiste em armazenar os dados no DW.

Devido às diferentes formas de coleta de dados utilizadas, por decisão de projeto, foi desenvolvido um script de extração para cada base de dados. Cada script vai direcionar os dados brutos para uma tabela flat, que vai armazenar cada um desses dados em uma tabela no mesmo formato dos dados de origem no servidor de banco de dados.

Para extrair os dados que são ofertados em formato CSV na base de dados do governo, como Notificações de Síndrome Respiratória Aguda Grave (SRAG) e Registro de Casos e Óbitos de COVID-19, foram desenvolvidos scripts[2] em Python. Esses scripts baixam os dados do site e carrega no programa o conteúdo da base de dados externa utilizando a linguagem Python. Também foi utilizada a biblioteca Pandas[3], que segundo McKinney (2017), faz com que seja mais fácil e rápido trabalhar com dados estruturados em Python, além de combinar processamento de alto desempenho com recursos flexíveis de manipulação de dados. Depois que os dados estão carregados na memória do programa, a biblioteca Pandas é utilizada para carregá-los, sem alteração, em uma tabela SQL do banco de dados Postgresql, a chamda tabela *flat*. Essa tabela será armazenada em uma Staging Area, qu, no contexto de DW é uma área de transição dos dados, onde eles são guardados para depois serem utilizados na transformação e carregamento. Dessa forma, evita-se trabalhar com os dados direto da base de dados de origem, melhorando a performance, pois evita que o programa tenha que ficar consultando diretamente da base de dados de origem (Santos, 2020).

Uma vez que os dados estão na Staging Area, os dados da tabela flat serão extraídos por outro script Python, que utiliza a biblioteca psycopg2 para integração com o banco de dados PostgreSQL, além da biblioteca Pandas. Com as bibliotecas psycopg2 e Pandas, os dados serão manipulados para serem transformados em um modelo dimensional, que facilita a análise de métricas com base em agregação. Uma vez que o script carregou os dados no programa e os transformou, a biblioteca Pandas se encarrega de carregar as novas tabelas no DW[4].

2.2. Data Warehouse

O banco de dados escolhido para guardar os dados do DW foi o PostgreSQL, que é um banco de dados relacional muito difundido no mercado, totalmente gratuito e open source, além da afinidade do desenvolvedor com a ferramenta. Os dados são estruturados em um modelo dimensional no esquema estrela (star schema[5]), que Piton (2017) descreve da seguinte forma:

> Star schema ou esquema estrela, idealizado por Ralph Kimball, é o modelo mais utilizado na modelagem dimensional para dar suporte à tomada de decisão e melhorar a performance de sistemas voltados para consulta. O esquema estrela é composto no centro por uma tabela fato, rodeada por tabelas dimensão, ficando parecido com a forma de uma estrela. A ideia é propor uma visão para modelagem de base de dados para sistemas de apoio à decisão, que é o caso do Data Warehouse.

[2] https://github.com/alberisrox/etl_framework.git
[3] https://pandas.pydata.org/
[4] https://www.devmedia.com.br/data-warehouse-ou-data-mart-por-onde-comecar/6996
[5] https://docs.microsoft.com/pt-br/power-bi/guidance/star-schema

2.3 Ferramenta de BI

O termo Business Intelligence (BI), aqui utilizado, é definido por Braghittoni (2017) como:

> [...]um conjunto de conceitos e métodos para melhorar o processo de tomada de decisão das empresas, utilizando-se de sistemas fundamentados em fatos e dimensões. O B.I. baseia-se em agrupar informações de diversas fontes e apresentá-las de forma unificada e sob uma métrica comum, a fim de que indicadores aparentemente distantes possam fazer sentido entre si.

Como ferramenta de BI, será utilizada a plataforma Metabase[6], que é uma ferramenta open source que permite a leitura dos dados do banco de dados, além de permitir o cruzamento das informações através de uma interface gráfica intuitiva e geração de gráficos para visualização em dashboards.

3. Modelagem de dados

3.1. Conjunto de dados

O foco central do projeto são os dados relacionados à pandemia do COVID-19 disponibilizados pelo governo. Inicialmente, esses dados serão extraídos de 2 fontes diferentes: notificações de Síndrome Gripal (SG), com o registro da testagem de casos leves; e Notificações de Síndrome Respiratória Aguda Grave (SRAG), com o registro de casos moderados e graves da doença. Essas bases de dados são oriundas do sistema OpenDatasus, desenvolvido pelo governo para disponibilizar dados públicos relacionados à saúde. Essas bases de dados contêm microdados para cada registro de caso de COVID-19 no nível do paciente, contendo informações como o município de residência do paciente infectado, idade, sexo, sintomas, tipo de teste realizado e data da notificação.

Muitos dos dados não estão presentes em apenas uma base de dados, mas em ambas as bases, dessa forma a heterogeneidade[7] dos dados tem que ser tratada, para que a base se torne o mais homogênea possível. Por exemplo, nas duas bases de dados, a idade é representada de formas diferentes, na base de dados SG, a idade é representada por valor um inteiro, já na base de dados SRAG, a idade é representada por uma tabela, pois, também representa a idade em dias ou em meses. Porém, não é necessário que haja duas tabelas que representam a idade na base de dados, visto isso, houve a junção da com inteiro idade da base de dados SG, com a tabela que representa a idade na base de dados SRAG. Essa situação é um bom exemplo de heterogeneidade estrutural, que é planejada na modelagem dos dados e é feita na fase de transformação, no processo de ETL. A imagem a imagem a seguir, representa as duas representações citadas acima.

[6] https://www.metabase.com/
[7] http://www.uel.br/cce/dc/wp-content/uploads/TCC_PEDRO_GARBIM.pdf

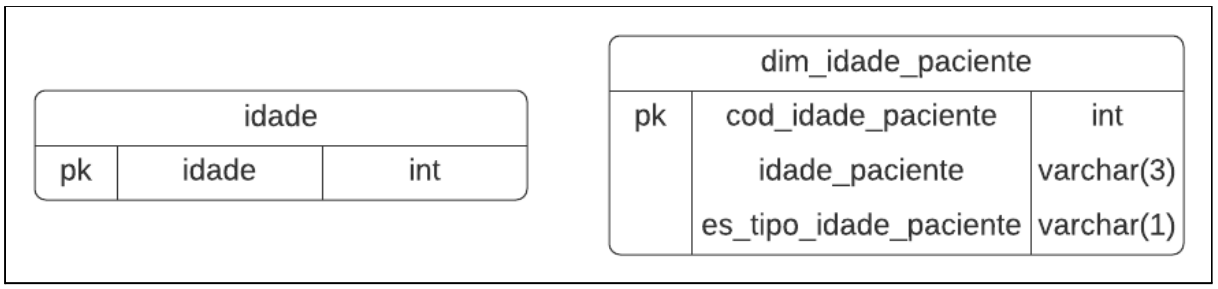


Figura 2. Ilustração da representação da idade nas bases SG e SRAG respectivamente.

Essa situação é tratada pelo processo ETL na fase de transformação dos dados. Ainda sobre o campo idade, nele também podemos identificar heterogeneidade sintática, visto que na base de dados sg o campo que representa a idade é nomeado “idade”, já na base de dados SRAG, apesar de não ser descrito o nome da tabela que representa a idade, os campos da tabela tem as seguintes representações: “NU_IDADE_N”, “TP_IDADE” e “COD_IDADE”. Depois de verificada essa situação, o campo no modelo dimensional passou a ser representado pela tabela “dim_idade_paciente”.

3.1.2. Notificações de Síndrome Gripal (SG)

Os dados de Síndrome Gripal são oriundos do sistema e-SUS NOTIFICA, disponibilizados a partir de março de 2020. Essa base contém dados referentes ao local de residência de pacientes, que apresentam sintomas gripais leves a moderados, mesmo que tenham sido notificados em outro local. Também contém outros dados sobre o paciente, como: sexo; idade; raça; os sintomas, como a descrição deles e a data de inicio; o tipo de teste utilizado e seu resultado; entre outras informações que fornecem uma visão geral das condições do paciente, do local onde ele reside e da fiscalização dos casos de covid-19.

3.1.3. Notificações de Síndrome Respiratória Aguda Grave (SRAG)

Apesar de o registro dos dados sobre a pandemia ter início em 2020, a vigilância de Síndrome Respiratória Aguda e Grave (SRAG) foi desenvolvida em 2009 para registro das notificações a respeito da pandemia da influenza A (H1N1). Em 2020, a vigilância da COVID-19 foi incorporada à rede de vigilância da influenza e outros vírus respiratórios.

Devido à incorporação da vigilância da COVID-19 em vigilâncias pré existentes, os dados disponibilizados por essa base contém variáveis desnecessárias para a plataforma. A relevância para a plataforma está apenas nos dados que se referem a pandemia da COVID-19.

A base de notificações de SRAG contém uma grande quantidade de atributos, cuja abrangência vai desde informações básicas do paciente, como sexo ou local de residência até informações como vacina tomada, comorbidades, se já foi internado anteriormente com os mesmos sintomas, entre outros dados. Grande parte desses dados serão descartados, restando apenas o que, de acordo com as decisões de projeto, forem relevantes. Como exemplo, temos o campo CS_ZONA que teria relevância para o projeto, por se tratar que um campo que divide o país em regiões específicas, porém foi notado uma incoerência nesses dados, onde uma mesma unidade de saúde tem registro como sendo cada uma das opções disponíveis em CS_ZONA.

A base de dados SRAG só notifica casos graves de síndrome respiratória, devido a isso, algumas métricas, como a taxa de óbitos, vai ser muito maior para a base SRAG do que para uma base que contenha notificações de casos leves e moderados.

3.2 Modelo dimensional

3.2.1. Modelo para Notificações de Síndrome Respiratória Aguda Grave (SRAG)

Para o banco de SRAG, um Star Schema foi modelado e implementado, contendo 1 tabela fato e 5 tabelas dimensões. Na figura 3 é possível verificar as tabelas, colunas e tipos do banco dimensional.

Além das tabelas dimensão e fato, foram incluídas métricas como a quantidade de internações hospitalares, internações em UTI, óbitos e notificações de casos de COVID-19.

Com a soma dessas variáveis, é possível mesclar uma boa quantidade de informações a respeito de pacientes com síndrome respiratória aguda grave.

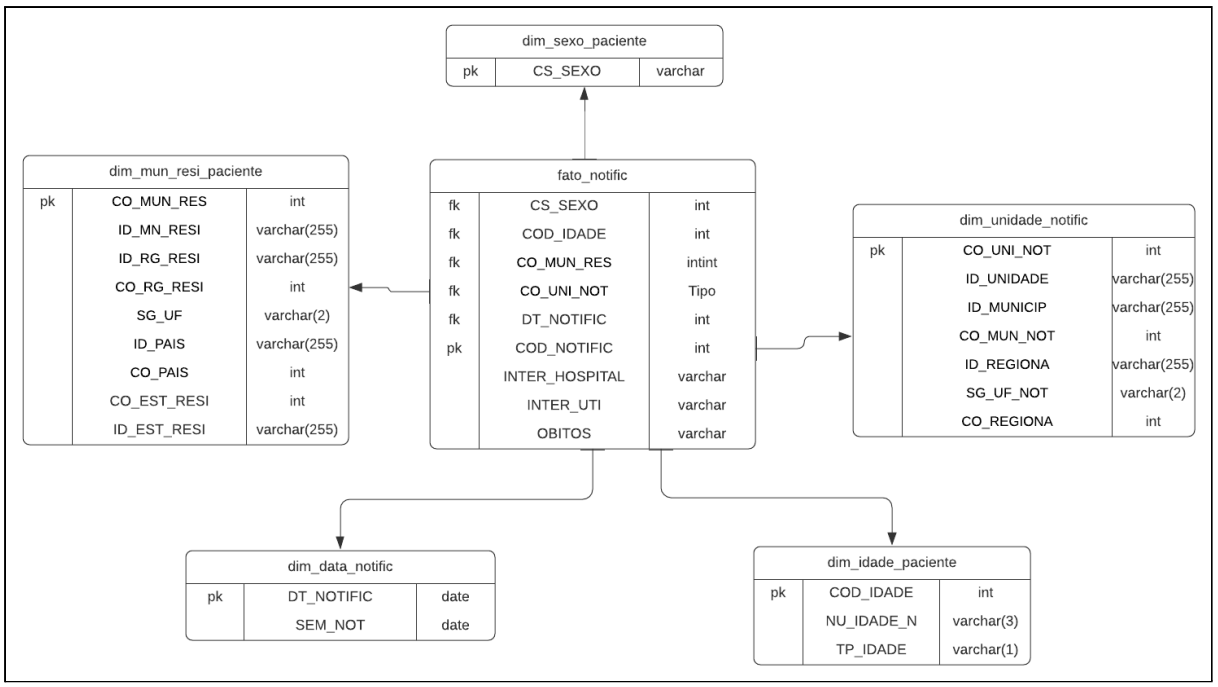
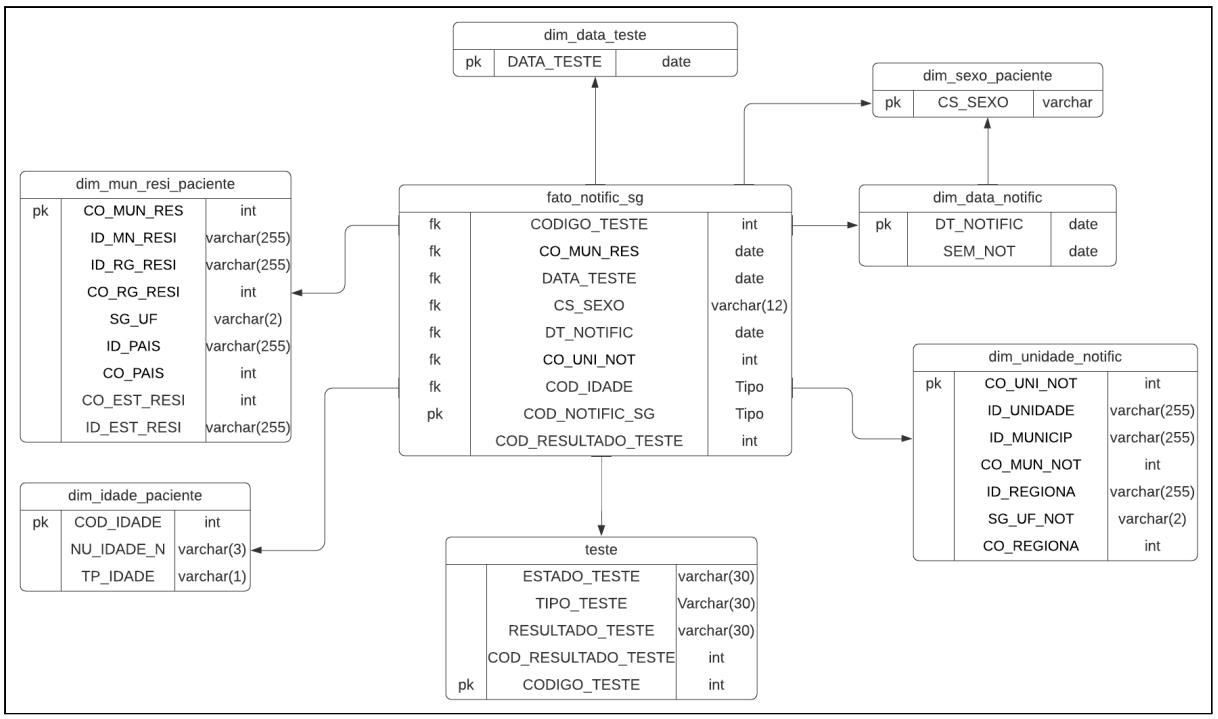


Figura 3. Ilustração do star schema SRAG.

3.2.2. Modelo para Notificações de Síndrome Gripal (SG)

Para o banco de SG, um Star Schema foi modelado e implementado, contendo 1 tabela fato e 7 tabelas dimensões. Na figura 4 é possível verificar as tabelas, colunas e tipos do banco dimensional.

Nessa tabela foi incluída a métrica “RESULTADO_TESTE”, que é o foco principal dessa tabela fato, trazendo a informação de número de casos positivos de COVID-19.



4. Solução

4.1 Solução ETL

A plataforma é dividida basicamente em 2 módulos, que são os componentes ETL e a plataforma de BI, Metabase.

O sistema ETL é simplificado, o código foi desenvolvido pensando em reduzir a participação do usuário no processo de extração e transformação. Quando o usuário desejar criar um modelo estrela, caberá a ele apenas as escolhas referentes à modelagem da tabela. Quando essa etapa for superada, o usuário precisará de um arquivo csv com informações básicas, como um nome para a tabela flat, nomes para as tabelas dimensão e fato e por último, o nome das colunas, isso organizado no seguinte modelo:

Tabela flat	Tabela Dimensão	Colunas
flat_exemplo	dim_exemplo	col_1,col_2,col_3

Tabela 1. Ilustração do preenchimento do arquivo csv.

Por fim, o usuário terá que informar o caminho para o csv acima informado e a fonte dos dados, seja um arquivo em seu computador ou um link para download, e o formato em que o dado se apresenta. Para facilitar a utilização dos componentes ETL por parte do usuário, a

plataforma disponibiliza uma interface gráfica simples, apenas para preenchimento das informações necessárias.

A figura 5 demonstra a interface que a plataforma utiliza para receber as informações:

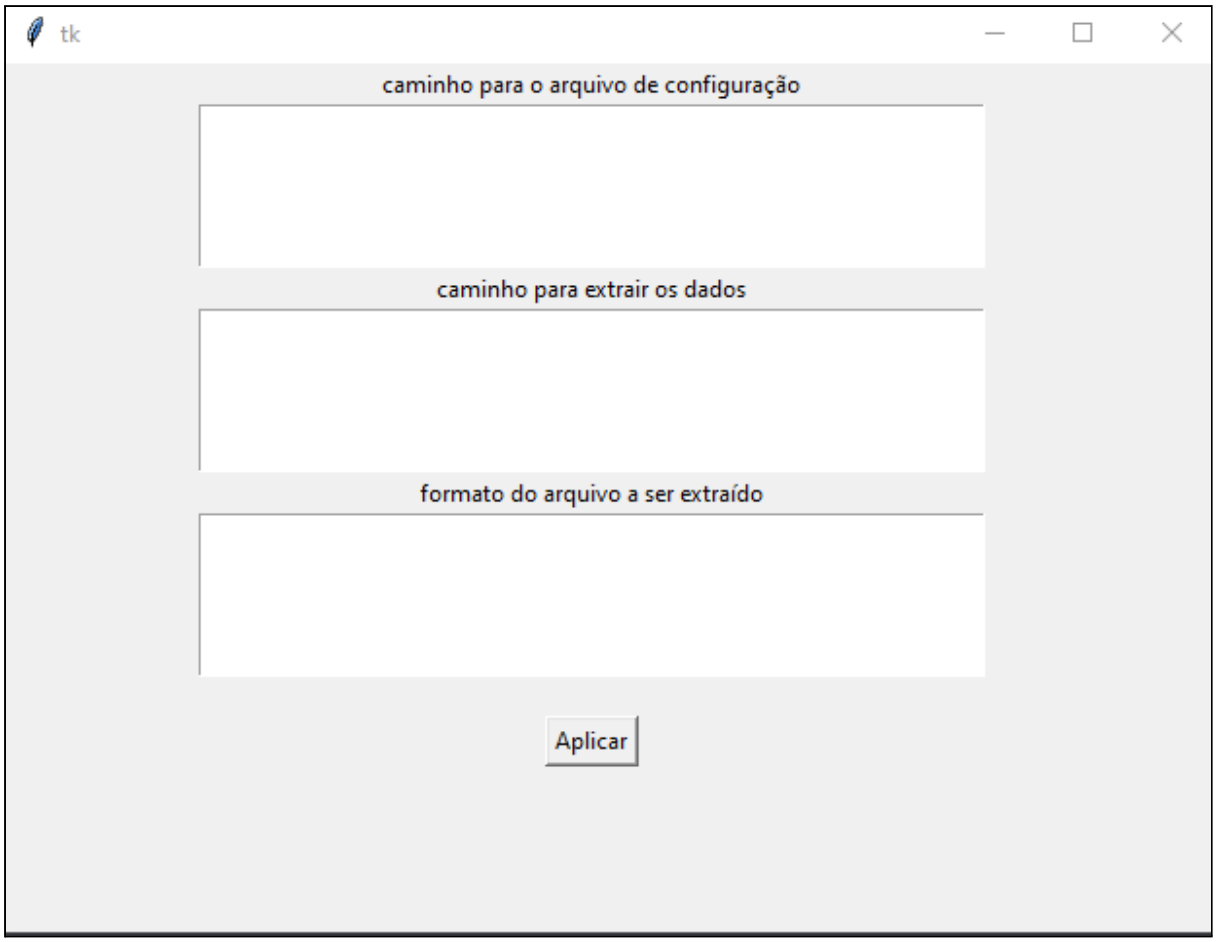


Figura 5. Interface para recebimento de informações necessárias para o processo ETL.

4.2 Plataforma de B.I

A plataforma Metabase disponibiliza diversos recursos para análise das informações do banco de dados. Uma vez criado o banco de dados, não é necessária uma configuração avançada para que a plataforma tenha acesso aos dados do banco e o usuário possa criar perguntas de negócio para a geração de gráficos e organização de dashboards.

A figura a seguir mostra uma consulta elaborada na própria plataforma, onde são exibidos os números de casos de SRAG por faixa etária.

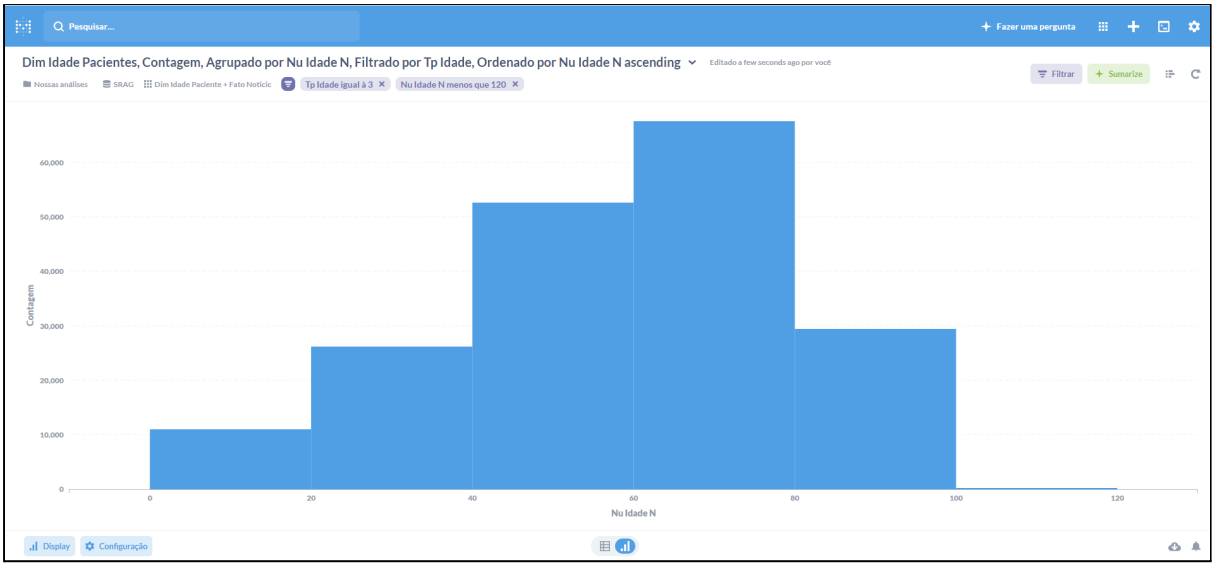


Figura 6. Número de notificações de Síndrome respiratória aguda grave.

5. Conclusão

Este projeto criou uma uma plataforma de Business Intelligence genérica que utiliza o processo ETL para extrair dados de bases de dados distintas, transformar os dados extraídos, com base em modelos baseados em bancos de dados e, somado ao Metabase, fornece uma plataforma analítica que permite a análise personalizada dos dados. Tendo em vista a estrutura inicial dos dados relacionados à COVID-19, especificados no documento, a ferramenta otimiza a análise desses dados, a ponto de tornar a análise deles acessível a pessoas que não tenham conhecimento de programação ou de bancos de dados, tornando a informação acessível a um número maior de pesquisadores.

Embora a ferramenta forneça o que foi proposto nesse documento, ainda há muitos avanços a serem realizados, para que a ferramenta se torne mais completa e facilite ainda mais a análise dos dados. Para atualizações futuras, é possível criar, previamente, novos painéis, com base no que é mais relevante para a pesquisa sobre a COVID-19, adicionar novas bases de dados à plataforma e realizar análise de dados com base nos dados armazenados no modelo dimensional.

Referências:

BRAGHITTONI, R. **Business intelligence**: implementar do jeito certo e a custo zero. São Paulo: Casa do Código, 2017.

BRAHIM, P. L. G. **Integração de Dados Heterogêneos em Big Data**: Estudo das Técnicas. Orientador: Prof. Dr. Vitor Valério de Souza Campos. 2021. 32. TCC (graduação) - Bacharelado em CIência da Computação, Universidade Estadual de Londrina, Londrina. 2021. Disponível

em: <http://www.uel.br/cce/dc/wp-content/uploads/TCC_PEDRO_GARBIM.pdf>. Acesso em: 03 de out. de 2021.

FERREIRA, J. et al. O Processo ETL em Sistemas Data Warehouse. **INForum 2010** - II Simpósio de Informática, Luís S. Barbosa, Miguel P. Correia (eds), 9-10 Setembro, 2010, pp. 757–765. Disponível em: <https://www.researchgate.net/publication/265195317_O_Processo_ETL_em_Sistemas_Data_Warehouse>. Acesso em: 02 out. 2021.

MELLIS, F. Com vacinação, nº de mortes por covid no Brasil cai 65% em 100 dias. **R7**. 2021. Disponível em: <https://noticias.r7.com/saude/com-vacinacao-n-de-mortes-por-covid-no-brasil-cai-65-em-100-dias-01092021> Acesso em: 03 de out. de 2021.

PITON, Rafael. **Data Warehouse – O Que É Star Schema?** Disponível em <https://rafaelpiton.com.br/blog/data-warehouse-star-schema/> Acesso em 02 out. 2021.

PRIMAK, F. V. **Decisões com B.I Business Intelligence**. 1. ed. Rio de Janeiro: Ciência Moderna Ltda., 2008

SANTOS, R.Z. Staging Area e ODS. **Medium,** 2020. Disponível em: <https://medium.com/@ricardozontasantos/staging-area-e-ods-c573c0d6477>. Acesso em: 03 de out. de 2021