Covid-19: vacinação global e seu impacto no índice de mortalidade

Projeto da Disciplina de BI

Prof. Anderson Nascimento
prof.anderson@ica.ele.puc-rio.br

Componentes do Projeto:
José Leonel Majewski – josemajewski@gmail.com
Lais Moreira dos Santos – matmslais@gmail.com
Márcio Del Rei – marcio.delrei@gmail.com

Sumário

1 Introdução

Este documento tem por finalidade descrever todo o processo de elaboração do projeto de BI aplicado ao estudo de caso ***"Covid-19: vacinação global e seu impacto no índice de mortalidade"***. Os principais objetivos desse estudo de caso são analisar a evolução temporal do número de óbitos antes e depois do início da campanha de vacinação e fornecer dados quantitativos e qualitativos sobre o número de vacinados e óbitos em cada país do mundo. Para tanto, nos baseamos em dados públicos que encontram-se disponíveis no site da revista digital *Our World In Data* (ver Seção 3).

A seguir, sumarizamos as principais etapas desenvolvidas ao longo desse projeto:

- criação de um JOB no PDI para automatizar a extração dos dados diários na referida base de dados.

- padronização e limpeza da base de dados extraída do site supracitado. Carga desses dados em uma Stage Area.

- construção do modelo multidimensional no Power Architect e do Data Warehouse (DW) no PostgreSQL.

- carga do DW com os dados tratados contidos na Stage Area.

- criação de um dashboard, que permita o usuário fazer uma análise exploratória dos dados sobre vacinação e mortalidade no mundo.

2 Estudo de Caso

2.1 Descrição do Estudo de Caso

A Covid-19 é uma doença infecciosa causada pelo vírus SARS-CoV-2, que foi identificado pela primeira em vez em humanos na cidade de Wuhan (China), em novembro de 2019. Desde então, o mundo sofre uma grave crise sanitária, que já totaliza 190 milhões de infectados e 4,2 milhões de óbitos. Diante desse cenário, houve um esforço mundial de cientistas e instituições em busca de vacinas que pudessem conter a propagação do vírus.

Em 8 de dezembro de 2020, pouco mais de um ano após o início da pandemia, o Reino Unido começou a vacinar sua população, tornando assim a vacina contra SARS-CoV-2 a mais rápida da história. Os reflexos da vacinação já puderam ser sentidos em alguns lugares do mundo. Na cidade de Serrana (SP-Brasil), por exemplo, onde toda a população adulta foi imunizada, o número de óbitos por Covid-19 teve uma queda de 95% e de internações 86% [1]. Embora a comprovação da eficácia da vacina nos deixe otimistas, por outro lado sabemos que a demanda global por doses torna a sua distribuição não equitativa, de modo que uma grande parcela da população mundial ainda não teve acesso à imunização. Além disso, a descoberta de novas variantes do vírus, tem deixado a comunidade científica preocupada quanto à eficácia das vacinas atualmente em circulação. Nesse contexto, através dos dados obtidos em domínio público (ver Seção 3) e fazendo uso de ferramentas de BI, buscamos estudar os *fatos vacinação* e *óbitos* com base nas d*imensões tempo* e *local* e a partir disto:

- extrair informações sobre o desempenho de cada país no processo de vacinação,
- comparar o desempenho dos países e continentes nesse processo,
- analisar a influência da vacinação na taxa de mortalidade por Covid-19.

Ao final desse projeto, espera-se que o usuário tenha acesso a um dashboard de fácil acesso, que o possibilite obter um panorama global sobre os índices de vacinação e mortalidade em função da Covid-19 no mundo.

[1] Dados fornecidos pelo Instituto Butantan no dia 31/05/2021.

3 Descrição do Modelo Transacional

Nesta seção, apresentamos uma descrição da fonte de dados utilizada neste projeto.

3.1 Fonte - Dados globais sobre a Covid-19

URL: https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv

O link acima, dá acesso ao arquivo “owid-covid-data.csv” disponibilizado pelo *Our World in Data* (https://ourworldindata.org), que é atualizado diariamente desde fevereiro de 2020. O arquivo contém uma coleção de 60 dados sobre a Covid-19 no mundo, além de informações gerais sobre cada país.

A seguir, fornecemos a descrição das colunas que efetivamente foram utilizadas neste projeto, a descrição completa da documentação dessa base de dados pode ser consultada no arquivo dicionario_dados constante na pasta anexos:

Coluna	Descrição
iso_code	ISO 3166-1 alpha-3 –códigos de três letras do país
continent	Continente
location	Localização geográfica
date	Data da observação
total_deaths	Total de mortes atribuídas à COVID-19
new_deaths	Novas mortes atribuídas à COVID-19
total_deaths_per_million	Total de mortes atribuídas à COVID-19 por 1.000.000 de pessoas
new_deaths_per_million	Novas mortes atribuídas à COVID-19 por 1.000.000 de pessoas
people_vaccinated	Número total de pessoas que receberam pelo menos uma dose de vacina

people_fully_vaccinated	Número total de pessoas que receberam todas as doses prescritas pelo protocolo de vacinação
people_vaccinated_per_hundred	Número total de pessoas que receberam pelo menos uma dose de vacina por 100 pessoas na população total
people_fully_vaccinated_per_hundred	Número total de pessoas que receberam todas as doses prescritas pelo protocolo de vacinação por 100 pessoas na população total
population	População em 2020

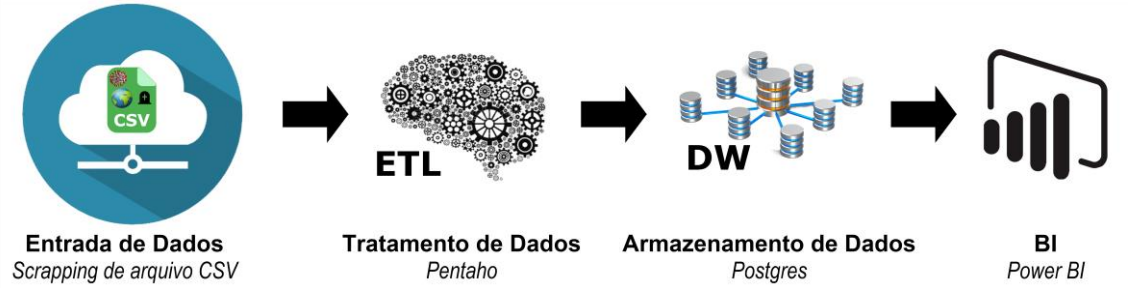
Nesta base de dados, nas colunas relacionadas aos dados sobre casos confirmados e mortalidade, podem ocorrer valores negativos. De acordo com o site *Our World in Data*, a razão para isto é a seguinte:

> "[…] O número de casos ou óbitos notificados por qualquer instituição num determinado dia - incluindo Johns Hopkins University(JHU), OMS, European Centre for Disease Prevention and Control (ECDC) e outros - não representa necessariamente o número real nessa data. Isso se deve à longa cadeia de notificação que existe entre um novo caso / óbito e sua inclusão nas estatísticas. Isso também significa que, às vezes, podem aparecer valores negativos em casos e óbitos quando um país corrige dados históricos, porque havia superestimado anteriormente o número de casos / óbitos. Alternativamente, grandes mudanças podem às vezes (embora raramente) ser feitas em toda a série temporal de um país se JHU decidir (e tiver acesso aos dados necessários) para corrigir os valores retrospectivamente."

4 Proposta de Processo de BI

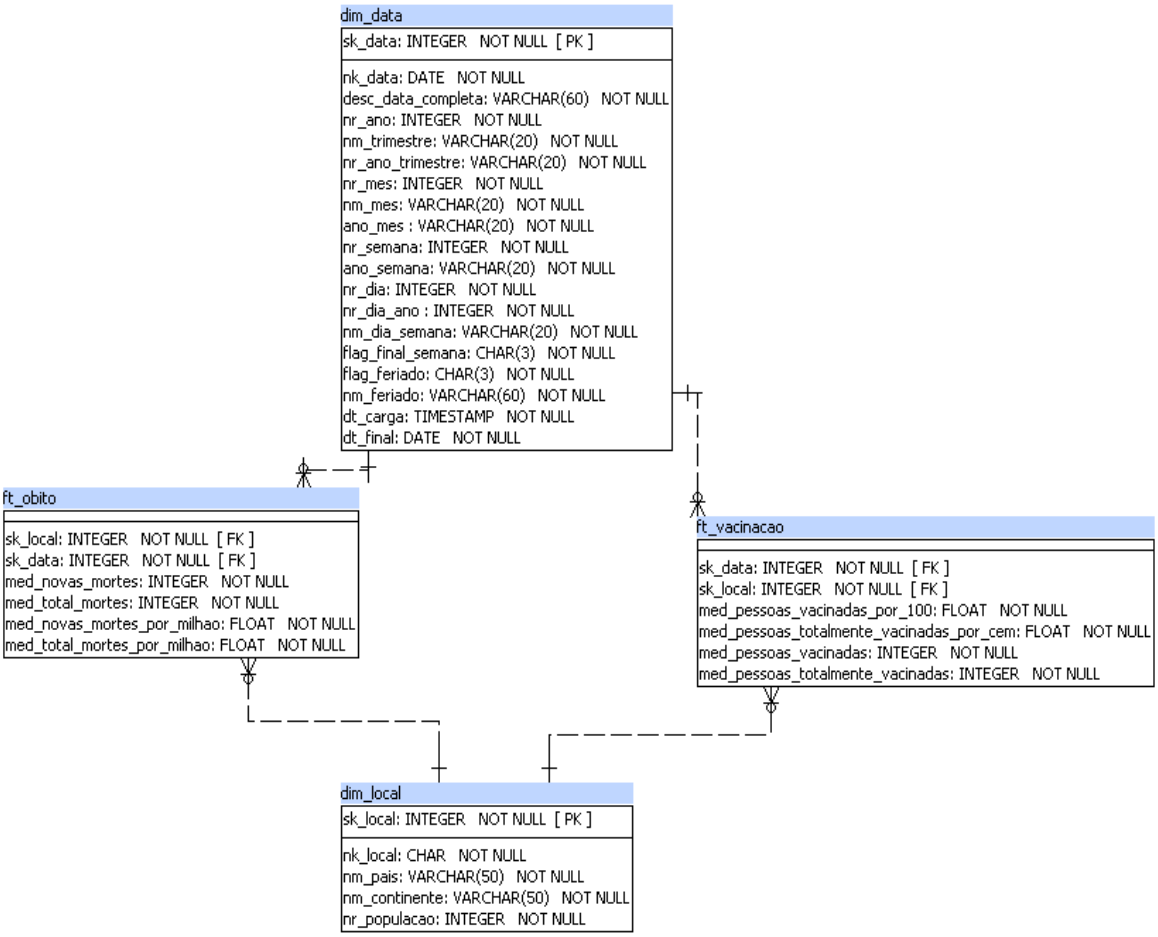
O processo proposto neste projeto, consiste inicialmente na criação de um JOB no software PDI, que rodará todos os dias às 23:50h realizando a coleta de um CSV, via Http Client - recurso da própria ferramenta, contendo informações sobre vacinações e mortalidade sobre a COVID-19 mundialmente. Em sua segunda fase, realizará o tratamento da base de dados "owid-covid-data.csv", mantida por Our World in Data (https://ourworldindata.org). Esta etapa consiste em utilizar a base de dados e a partir dela começar o processo de ETL gerando assim de forma automatizada o DW que será consumido pelos Dashboards desenvolvidos e apresentados no Power BI. Estes Dashboards apresentarão informações relevantes para indicar a eficácia das vacinas em contraste com o número de óbitos de forma mundialmente macro, mostrando visibilidade, transparência e apoio a tomadas de decisões com informações atualizadas até o dia anterior (D-1) ou sob demanda.

O fluxo completo das etapas envolvidas neste projeto, desde a coleta de dados até a criação dos dashboards, é representada pelo croqui a seguir:



5 Modelo Multidimensional

Esta seção apresenta o modelo dimensional do estudo de caso, que é composto por duas tabelas de fatos *ft_vacinacao* e *ft_obitos*, onde cada uma delas se corresponde com as dimensões conformadas *dim_data* e *dim_local.* O diagrama do modelo dimensional é dado a seguir:



Este modelo foi construído no SQL Power Architect 1.0.9 e a partir dele as tabelas de fatos e dimensões, que compõem o data warehouse, foram geradas no PostgreSQL (pdAdmin 4.30).

A tabela fato *ft_obito* é composta pelas *surrogate keys* das dimensões *dim_local* e *dim_data* (que em *ft_obito* desempenham o papel de *foreign keys*), pelas métricas aditivas *med_novas_mortes, med_novas_mortes_por_milhao* e pelas métricas semi-aditivas *med_total_mortes* e *med_total_mortes_por_milhao*. As métricas *med_novas_mortes e med_novas_mortes_por_milhao* representam as novas mortes e as novas mortes por 1.000.000 de pessoas atribuídas à COVID-19, respectivamente.

A tabela fato *ft_vacinacao* contém dados sobre os índices de vacinação no mundo e é composta pelas *surrogate keys* das dimensões *dim_local* e *dim_data* (que em *ft_vacinacao* desempenham o papel de *foreign keys*) e pelas métricas semi-aditivas *med_pessoas_vacinadas, med_pessoas_vacinadas_por_100, med_pessoas_totalmente_vacinadas* e

med_pessoas_totalmente_vacinadas/1000. A descrição de cadas uma dessas métricas é dada a seguir:

- ***med_pessoas_vacinadas:*** número total de pessoas que receberam pelo menos uma dose da vacina. Se uma pessoa receber a primeira dose de uma vacina de 2 doses, essa métrica aumenta em 1. Se ela receber a segunda dose, a métrica permanece a mesma.

- ***med_pessoas_vacinadas_por_100:*** *med_pessoas_vacinadas* por 100 pessoas na população total do país.

- ***med_pessoas_totalmente_vacinadas:*** número total de pessoas que receberam todas as doses prescritas pelo protocolo de vacinação. Se uma pessoa recebe a primeira dose de uma vacina de 2 doses, essa métrica permanece a mesma. Se ela recebe a segunda dose, a métrica sobe 1 unidade.
- ***med_pessoas_totalmente_vacinadas_por_100:*** *med_pessoas_totalmente_vacinadas* por 100 pessoas na população total do país.

A dimensão *dim_data* foi extraída do script fornecido pelo professor e contém um detalhamento completo das datas de notificações de óbitos e vacinação.

Por fim, a dimensão *dim_local* comporta dados sobre cada país, como nome (*nm_pais*), nome do continente onde o país em questão se localiza (*nm_continente*) e número de indivíduos na população (*nr_populacao*).

Observe que o modelo dimensional adotado aqui é de alta granularidade, já que os dados fornecem pouco detalhamento sobre os fatos considerados.

6 Elaboração do Data Warehouse

O Data Warehouse é a fonte integradora de informações, utilizada com o intuito de servir de base para a camada de aplicação que é responsável por fornecer dados para os dashboards, instrumento este que será utilizado pelo público para extrair informações sobre vacinação e mortalidade em função da Covid-19.

6.1 Definição do DW

6.1.1 Arquitetura

Neste projeto, o Data Warehouse tem uma arquitetura independente. De fato, os dados extraídos da base de dados, após serem tratados, compõem um único banco de dados que atende a necessidade específica do negócio.

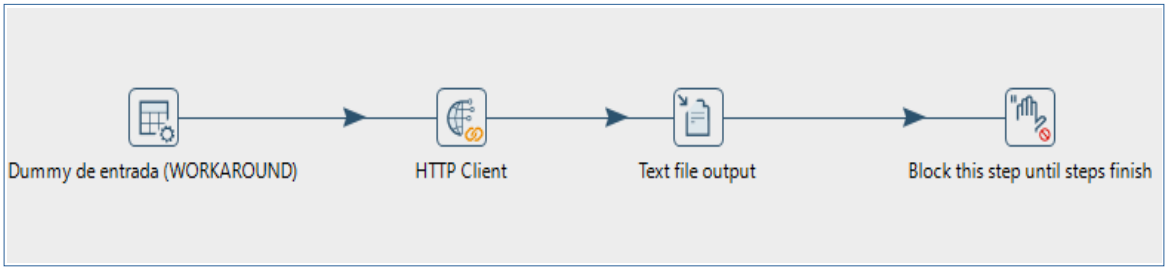
6.1.2 Arquitetura Física

A arquitetura física adotada é do tipo on-premises, uma vez que os dados foram armazenados no computador pessoal. Mais precisamente, a pasta C:\Projeto_BI foi criada no Disco C para armazenar todas as transformações e o JOB do PDI (Pentaho), responsáveis pelo tratamento e automatização do processo de ETL e alimentação do DW dinamicamente. Diariamente este JOB é acionado através de um agendador de tarefas do Windows. Através desse artifício, evita-se que o usuário tenha que estar logado no PDI no instante da atualização, basta que o computador esteja conectado à internet no instante programado, à escolha do usuário. Com isso, todos os dias o schema dw_covid19 criado no database projeto_covid19 e construído no SGBD pgAdmin 4 (local), contendo os dados dos fatos e dimensões, é automaticamente atualizado.
A escolha de utilizar esta arquitetura física, deve-se ao fato de que o Disco C é um ambiente universal, de modo que os usuários que desejarem executar este projeto não terão que alterar nenhum diretório nas transformações e no JOB. No entanto, esta solução é adaptável para qualquer ambiente, seja ela on-premise ou em nuvem.

7 Projeto de ETL

7.1 Descrição do Projeto de ETL

O ETL foi realizado utilizando a versão 9.1 do Pentaho PDI.
Na etapa de *extração* (veja arquivo *00_Extracao_CSV_Github*), foi criado uma transformação no software PDI responsável por coletar da fonte https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv o arquivo .csv fornecido pelo *Our World in Data* (veja seção 3) e depositá-lo no diretório C:\Projeto_BI\stage.



Na etapa de *transformação* (ver arquivo 01_ETL1_Pre_DW), os seguintes steps foram executados:

- Leitura da base de dados “owid-covid-data.csv” através da transformação CSV Input.

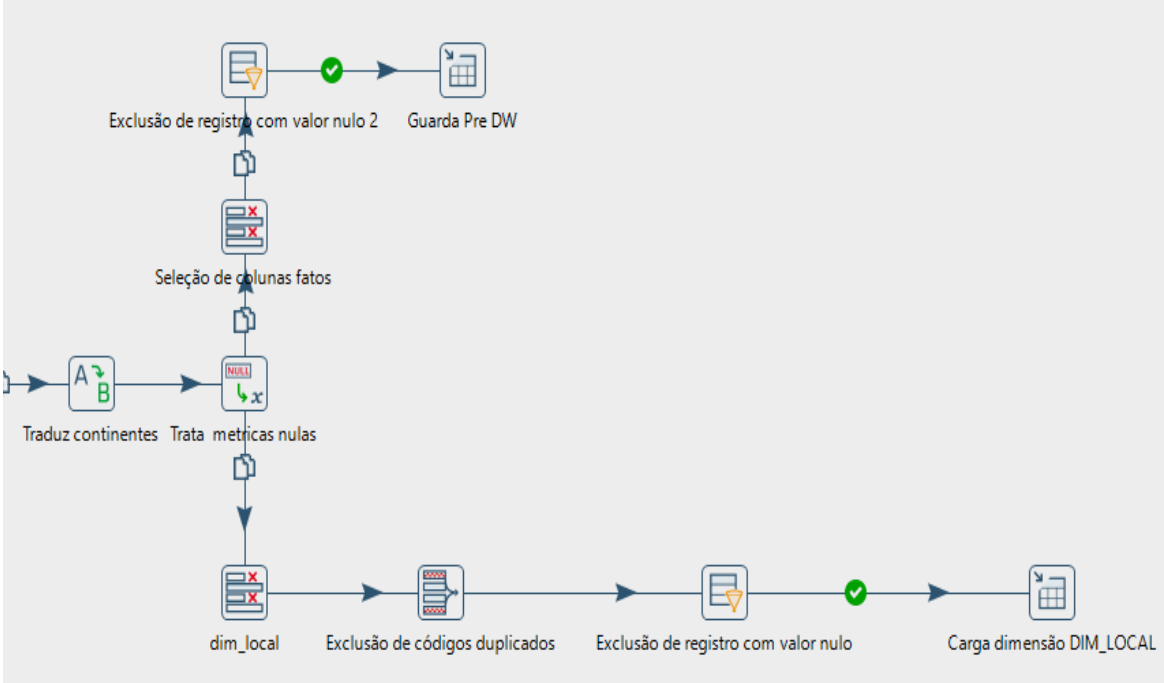
- Seleção de colunas úteis por meio da transformação Select/Rename Values.

- Eliminação das linhas contendo dados referentes à continentes e mundial por meio do step Filter Rows. Este procedimento se deve ao fato de que na base de dados, além de serem fornecidos os dados de cada país, também são fornecidos dados por continente e dados mundiais. Entretanto, esses últimos se tornam redundantes, já que podem ser obtidos através de um simples agrupamento. Para evitar tal redundância, eliminamos essas linhas da base de dados.

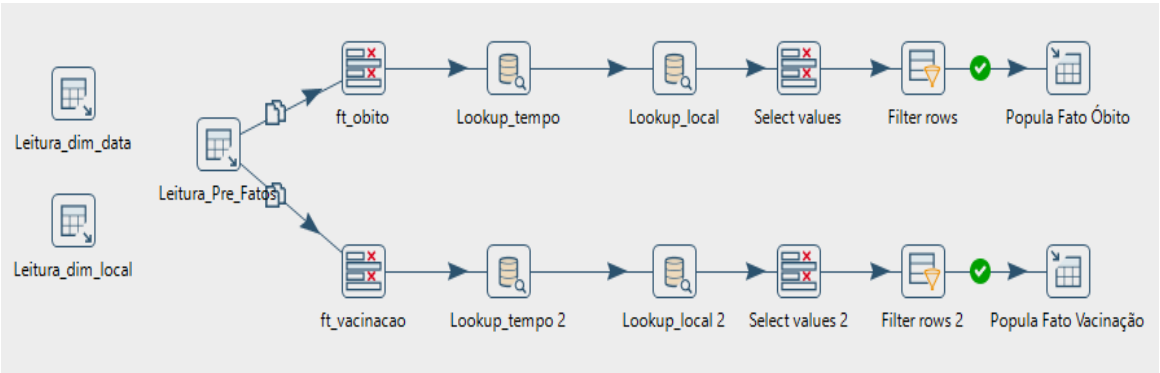
- Alteração no nome dos atributos através da transformação Select/Rename values, para atender ao design pattern.

- Tradução dos nomes de países e continentes, que encontravam-se em inglês na base de dados original. Esse procedimento foi executado por meio de mapeamento usando o step Value Mapper.

- Tratamento de valores faltantes nos campos med_total_mortes, med_novas_mortes, med_total_mortes_por_milhao, med_novas_mortes_por_milhao, med_pessoas_vacinadas, med_pessoas_totalmente_vacinadas, med_pessoas_vacinadas_por_100 e med_pessoas_totalmente_vacinadas_por_cem. Os valores faltantes foram substituídos por zero através da transformação Replace Null Value.

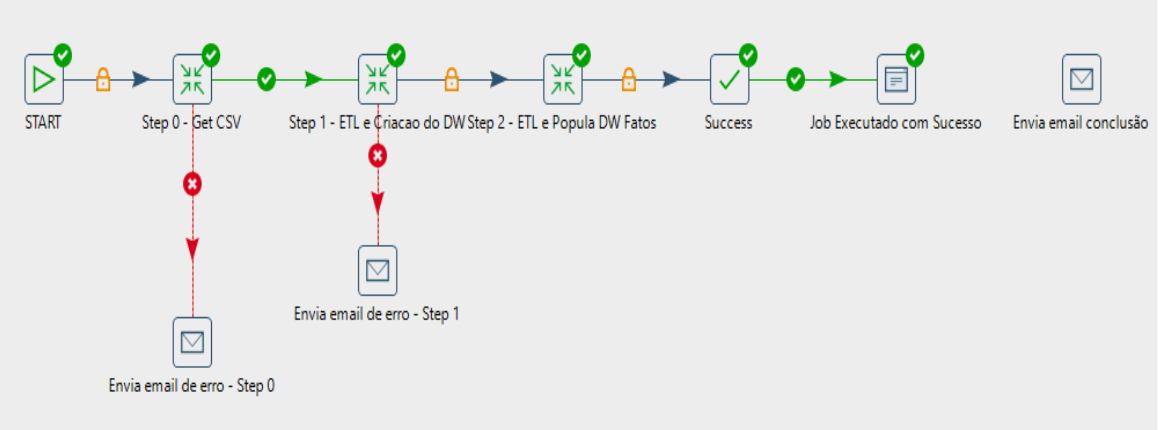


Na etapa de *carga* do DW, inicialmente criamos as estruturas das tabelas dim_data, dim_local, ft_obito e ft_vacinacao através do SQL gerado no Power Arquitect e do script para dim_data fornecido pelo professor. A carga dos dados na tabela dim_local e do banco de apoio pre_dw_covid19 foi feita no final do fluxo apresentado acima. O banco de dados pre_dw_covid19, foi usado apenas como base temporária das transformações para carga das tabelas fato.
A carga das tabelas ft_obito e ft_vacinacao foi feita em uma terceira transformação, cujo fluxo é apresentado a seguir:



Finalmente, para automatizar o processo de ETL, foi criado um JOB (ver arquivo Covid19_job.kjb). Optamos por dispará-lo diariamente por meio de um agendador de tarefas do Windows. Uma vez acionado, o JOB executa as três

transformações mencionadas anteriormente, fornecendo como resultado final a versão mais atual dos dados no Data Warehouse. O fluxo do JOB é o seguinte:



8 Dashboard

8.1 Descrição da Elaboração

O dashboard foi elaborado no Power BI (versão julho/2021) com a proposta de fornecer ao usuário uma visualização das taxas de mortalidade e vacinação por país, além de permitir que sejam feitas análises comparativas sobre o desempenho dos países no processo vacinação e sobre a evolução do número de óbitos.

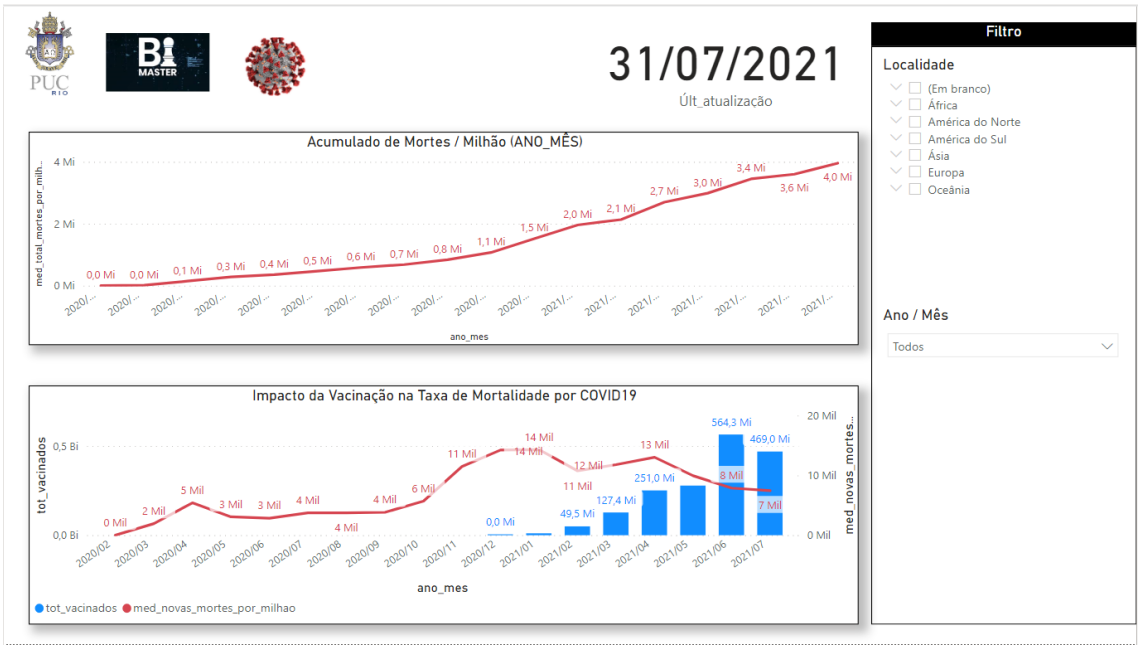
8.2 Telas do Dashboard

Em todas as telas do dashboard, os dados podem ser filtrados por país, continente e período. Na primeira tela do dashboard, o usuário pode visualizar:

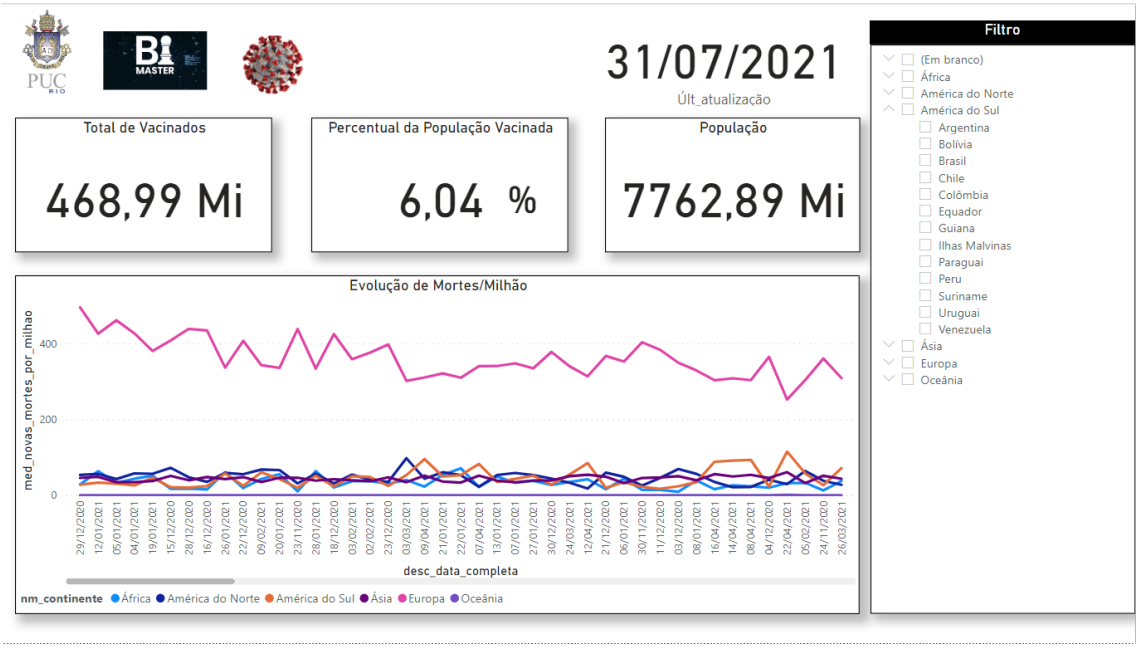
1) a data da última atualização dos dados

2) um gráfico de linhas do acumulado de mortes por milhão em função do tempo.

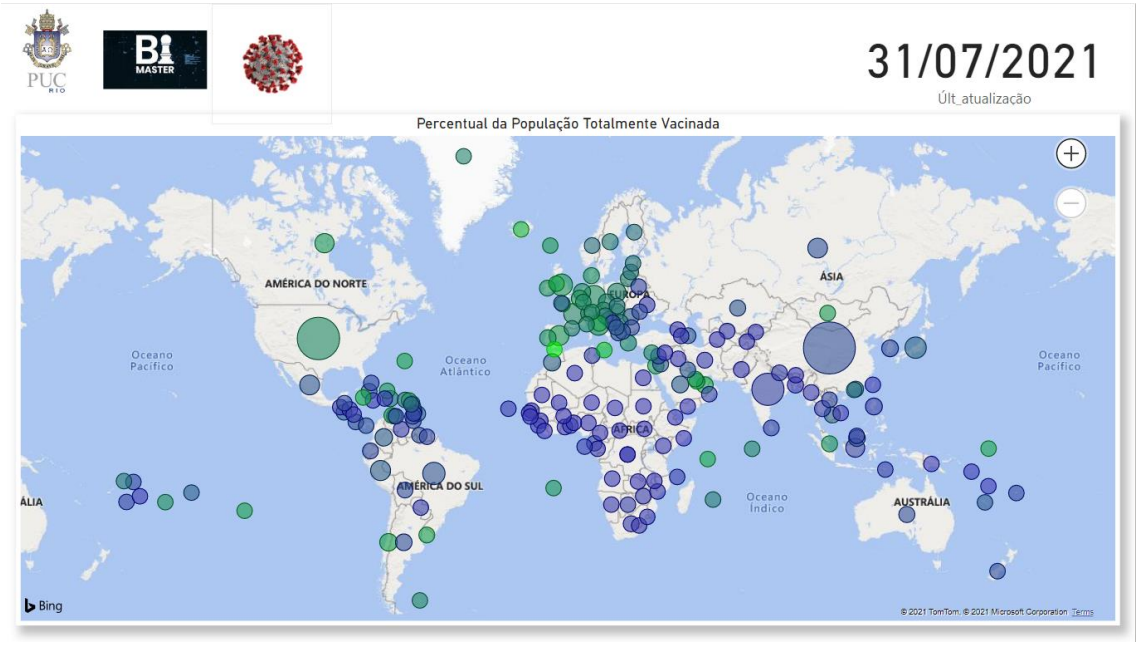
3) um gráfico de colunas agrupadas e linha em função do tempo. A curva de contínua retrata o número de novos óbitos por milhão, enquanto no gráfico de barra são representados o total de vacinas aplicadas no período.



Na segunda tela do dashboard, o usuário te acesso aos cartões contendo o número total de vacinados, o percentual da população vacinada e a população atual. Além disso, é apresentado um gráfico retratando a evolução temporal do número de óbitos por milhão de habitantes.



A terceira tela é dedicada a representar o percentual de população totalmente vacinada. Além do mapa, apresenta a última data de atualização dos dados.



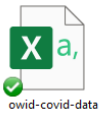
Finalmente, na última tela é retratado o número de óbitos por milhão até a última data de atualização.

9 Conclusão

No processo de desenvolvimento deste projeto, pudemos aplicar todas as etapas do processo de BI, permitindo a obtenção de informações em tempo real e a apresentação dos dados de forma a identificar graficamente a evolução do número de óbitos e o impacto da vacinação na redução da mortalidade decorrente da COVID-19. Para fornecer uma análise detalhada, o dashboard permite filtragem temporal, por continente e por país.

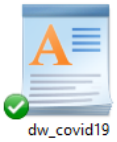
10 Anexos

Os arquivos utilizados e gerados neste projeto, estão na pasta da estrutura do projeto, citada na próxima seção. A descrição de cada um deles é dada a seguir:



Base de dados extraída do *Our World in Data.*

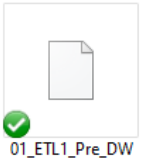
Dicionário dos dados contidos na base de dados “owid- covid-data”



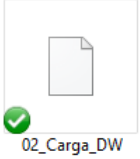
Arquivo no formato SQL contendo o script de criação das tabelas dim_data, dim_local, ft_obito, ft_vacinacao.

Modelo muldimensional gerado no SQL Power Architect 1.0.9

Arquivo ktr responsável pela extração da base de dados diretamente do site do *Our World in Data.*

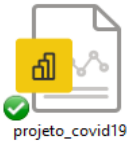


Arquivo ktr contendo todos os steps de transformação da base de dados.

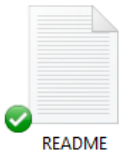


Arquivo ktr que executa a carga do DW.

Arquivo kjb que realiza a automatização do ETL.



Arquivo pbix contendo o dashboard.



Este arquivo explica como executar o projeto ponta a ponta, assim como os softwares e suas respectivas versões para instalação.

11 Arquivos

Estrutura dos arquivos do projeto:

```
C:\Projeto_BI
        |
        +---anexos
        |   |   README.txt
        |   |
        |   +---dicionario_dados_stage_owid
        |   |       dicionario_dados.pdf
        |   |       owid-covid-data_TEMPLATE.csv
        |   |
        |   +---modelo_dimensional
        |   |       Modelo_dimensional.architect
        |   |
        |   +---schedule_windows
        |   |       ProjetoBI_COVID19 (TaskWindows).xml
        |   |
        |   \---script_bd
        |           dw_covid19.sql
        |
        +---bi
        |       projeto_covid19.pbix
        |
        +---ktr
        |       00_Extracao_CSV_Github.ktr
        |       01_ETL1_Pre_DW.ktr
        |       02_Carga_DW.ktr
        |       Covid19_job.kjb
        |       job.bat
        |       log.txt
        |
        +---logs
        |       covid19.log
        |
        \---stage
                owid-covid-data.csv
```