COVID-19 – Análise de Indicadores de Saúde e Evolução da Pandemia

Bernardo Martins - aluno nº25436
Instituto Politécnico do Cávado e Ave
LESI - Integração de Sistemas de Informação
Prof.Óscar Ribeiro
16/10/2025

Indíce

Introdução..... 3
Arquitetura e Ferramentas..... 4
Processo ETL..... 5
Fase 1: Estruturação da Base de Dados..... 5
Fase 2: Extração e Carregamento dos Dados (Extract & Load)..... 5
Fase 3: Transformação e Enriquecimento dos Dados (Transform)..... 6
Fase 4: Agregação e Criação da Tabela Analítica..... 6
Fase 5: Análise e Criação de Indicadores Derivados..... 7
Visualização em Node-RED..... 8
Resultados e Conclusão..... 9
Referências..... 10

Introdução

O presente trabalho foi desenvolvido no âmbito da Unidade Curricular de Integração de Sistemas de Informação (ISI) e foca-se na aplicação de processos de ETL (*Extract, Transform, Load*) para analisar dados da pandemia de COVID-19. O objetivo central é integrar fontes de dados heterogéneas para calcular, analisar e visualizar métricas de saúde comparáveis entre diferentes nações, como o número de casos e mortes por 100 mil habitantes.

Para tal, foram utilizadas ferramentas de ETL, uma base de dados relacional para armazenamento e uma plataforma de visualização para a criação de dashboards interativos, cumprindo os requisitos propostos no enunciado do projeto.

Arquitetura e Ferramentas

A arquitetura tecnológica da solução foi concebida para ser robusta e escalável, integrando diversas ferramentas, sendo estas:

- **Base de Dados (MySQL 8.0):** Esta base de dados armazena de forma estruturada todos os dados brutos e processados, incluindo informações diárias de casos e mortes por COVID-19, dados populacionais e tabelas de mapeamento. A sua configuração foi otimizada para suportar consultas complexas e um volume crescente de informações, garantindo a integridade e disponibilidade dos dados para as camadas superiores da arquitetura.
- **Visualização e Integração (Node-RED):** Utilizado para a criação de dashboards interativos e intuitivos, permite a visualização dinâmica dos resultados das análises. Além disso, o Node-RED é responsável por orquestrar os fluxos de dados, conectando a base de dados a diferentes componentes de visualização e garantindo que os gráficos e relatórios sejam atualizados em tempo real ou com a frequência desejada.
- **Scripts Auxiliares (Python 3.14.0):** Para a manipulação, transformação e exportação de dados, foi utilizada a linguagem de programação Python. Os scripts desenvolvidos em Python são essenciais para:
    - **Processos ETL (Extract, Transform, Load):** Automatizam a extração de dados das fontes, a sua transformação para um formato adequado e o carregamento para a base de dados MySQL. Isso inclui a limpeza, padronização e enriquecimento dos dados.
    - **Exportação de Análises Específicas:** Permitem a exportação automatizada de conjuntos de dados ou resultados de análises.
- **Fontes de Dados (CSV):** A solução integra dados provenientes de diversas fontes, todas em formato CSV, sendo as principais:
    - **cases_deaths.csv:** Este ficheiro contém o registo diário detalhado de casos confirmados e mortes relacionadas ao COVID-19.
    - **population_totals.csv:** Contém dados populacionais anuais, que são cruciais para a normalização de indicadores (por exemplo, casos por 100.000 habitantes) e para análises per capita. Estes dados são provenientes do World Bank, garantindo a sua fiabilidade e abrangência global.
    - **iso_country_codes.csv:** Uma tabela de mapeamento que fornece códigos de países nos formatos ISO Alpha-2 e Alpha-3. Este ficheiro permite que as análises sejam facilmente agregadas e visualizadas por país.

A combinação destas tecnologias cria um ecossistema eficiente para a gestão e análise de dados, capaz de fornecer *insights* valiosos a partir de fontes de dados diversas e complexas.

Processo ETL

O processo foi dividido numa sequência de cinco fases lógicas e modulares, cada uma implementada através de scripts SQL ou Python dedicados. Esta abordagem garante a reprodutibilidade, manutenção e clareza de todo o fluxo de trabalho, desde a criação da estrutura da base de dados até à análise final.

Fase 1: Estruturação da Base de Dados

Script: 01_create_tables.sql

A primeira fase consistiu na preparação do ambiente da base de dados ***health_data***. Este script inicial cria as tabelas "base" que servem como repositório para os dados brutos extraídos dos ficheiros CSV.

- Tabelas Criadas:
    - **iso_country_codes:** Armazena os códigos de referência dos países (ISO Alpha-2 e Alpha-3).
    - **population:** Contém os dados anuais de população por país.
    - **cases_deaths**: A tabela principal para os registos diários da pandemia. Foi criada já com a coluna *country_iso* (inicialmente nula) para ser preenchida na fase de transformação.

Fase 2: Extração e Carregamento dos Dados (Extract & Load)

Script: 02_load_data.py

Nesta fase, os dados são extraídos dos ficheiros CSV de origem e carregados para a base de dados. Para garantir um processo robusto e eficiente, esta etapa foi automatizada com um script Python.

- O script recorre às bibliotecas pandas para a leitura e manipulação dos dados e SQLAlchemy para a ligação e escrita na base de dados MySQL.
- **Processo de Carregamento:**
    1. O script inicia com a limpeza das tabelas *(TRUNCATE TABLE)*, assegurando que cada execução é feita do zero, evitando a duplicação de dados.
    2. Para o ficheiro de grande volume *(cases_deaths.csv)*, a leitura é realizada em chunks (“partes” de 50.000 linhas), uma técnica que otimiza o uso da memória e permite o processamento de grandes volumes de dados de forma eficiente.
    3. Os dados são inseridos nas tabelas existentes *(append)*, preservando a estrutura previamente definida.

Fase 3: Transformação e Enriquecimento dos Dados (Transform)

Script: 03_transform.sql

Com os dados brutos já carregados, esta fase foca-se na limpeza, normalização e enriquecimento da informação.

- **Normalização de Países:** O principal objetivo deste script é preencher a coluna *country_iso* na tabela *cases_deaths*. Isto é alcançado através de uma operação *UPDATE* com *JOIN* que cruza o nome do país com a tabela de referência *iso_country_codes*.
- **Controlo de Qualidade:** O script inclui uma consulta de verificação *(SELECT DISTINCT)* para identificar quaisquer países que não foram mapeados, permitindo a sua correção manual.

Fase 4: Agregação e Criação da Tabela Analítica

Script: 04_covid_annual_metrics.sql

Nesta fase, os dados diários detalhados são transformados numa tabela agregada e otimizada para análise e visualização.

- **Criação da Tabela:** O script cria a tabela *covid_annual_metrics*, que servirá de base para os dashboards e análises finais.
- **Processo de Agregação:**
    1. Os dados de casos e mortes diários são somados *(SUM)* para obter totais anuais por país *(GROUP BY)*.
    2. Estes resultados são cruzados *(JOIN)* com os dados de população correspondentes.
    3. As métricas chave, como cases_per_100k e deaths_per_100k, são calculadas.
- **Resultado:** É gerada uma tabela de factos consolidada, que permite consultas rápidas e eficientes.

Fase 5: Análise e Criação de Indicadores Derivados

Scripts: 05_add_mortality_rate.sql, 06_covid_total_deaths.sql, 07_covid_top10_mortality.sql

A fase final do processo consiste na criação de novas métricas e tabelas de resumo para análises mais aprofundadas.

- **Cálculo da Taxa de Mortalidade:** O script ***05_add_mortality_rate.sql*** adiciona a coluna *mortality_rate* à tabela *covid_annual_metrics* e calcula a percentagem de mortes em relação ao total de casos.
- **Criação de Tabelas de Sumário:**
    - ***06_covid_total_deaths.sql:*** Gera uma nova tabela com o total de mortes por país ao longo de todo o período.
    - ***07_covid_top10_mortality.sql:*** Cria uma tabela de ranking com os 10 países que apresentaram a maior taxa de mortalidade média, um indicador pronto para ser utilizado em relatórios e visualizações.

Visualização em Node-RED

A componente de visualização foi implementada em Node-RED, utilizando dois fluxos distintos para criar dashboards interativos.

- **Fluxo 1:** Dashboards de Casos e Mortes por País *(node-red-flow.json)* Este fluxo realiza uma única consulta à base de dados para extrair todos os dados da tabela covid_annual_metrics. A informação é depois distribuída por três nós Function, que filtram os dados para Portugal, França e Alemanha, preparando-os para serem visualizados em gráficos de linhas com a biblioteca *Chart.js*. Uma funcionalidade notável é o uso de dois eixos Y distintos para comparar "Casos/100k" e "Mortes/100k", mesmo com escalas de valores diferentes.
- **Fluxo 2:** Dashboards de Análise de Mortalidade *(node-red-mortality-rate.json)* Este fluxo cria um novo separador no dashboard ("Estatísticas") com duas visualizações:
    1. Um gráfico de linhas comparativo que mostra a evolução da taxa de mortalidade para os três países em foco.
    2. Uma tabela HTML dinâmica que apresenta o ranking da tabela *covid_top10_mortality*, mostrando os 10 países com a maior taxa de mortalidade média.

Resultados e Conclusão

A análise dos dados e dos dashboards gerados permitiu extrair várias observações relevantes. Os anos de 2020 e 2021 concentram a maioria das mortes globais, representando o período mais crítico da pandemia. Os gráficos interativos foram essenciais para comparar tendências anuais e perceber a evolução do impacto do vírus em cada país. A partir de 2021, por exemplo, nota-se que o COVID-19 tornou-se mais agressivo em Portugal, em comparação com França e Alemanha.

Uma limitação identificada nos dados foi a de que vários países, incluindo França e Alemanha, aparentam ter descontinuado o registo sistemático de casos e mortes em 2024. Adicionalmente, a análise da tabela do top 10 de mortalidade média revelou dados importantes: no caso do Iémen, mesmo com um número relativamente baixo de casos (menos de 15.000 entre 2020 e 2022), a taxa de mortalidade foi próxima de 20%, indicando uma letalidade muito elevada entre os casos registados.

Em suma, este projeto cumpriu com sucesso os objetivos propostos. Foi implementado um processo ETL completo, desde a extração de dados brutos com Python, passando pela sua transformação e agregação em MySQL, até à criação de dashboards interativos e informativos em Node-RED. A combinação das ferramentas demonstrou ser uma solução eficaz e robusta para resolver problemas de integração e análise de dados em cenários reais. Como trabalho futuro, o projeto poderia ser expandido para incluir outros indicadores, como dados de vacinação ou taxas de hospitalização, enriquecendo ainda mais a análise.

Gravação do Projeto Completo

Referências

Para a realização deste projeto, foram utilizadas as seguintes fontes de dados públicas:

1. **Códigos de Países (ISO Alpha-2 e Alpha-3):** O ficheiro iso_country_codes.csv foi obtido a partir do dataset "Country Codes Alpha2, Alpha3" disponível na plataforma Kaggle.
    - **URL:** https://www.kaggle.com/datasets/emolodov/country-codes-alpha2-alpha3
2. **Dados de Casos e Mortes por COVID-19:** A principal fonte de dados sobre a pandemia, *cases_deaths.csv*, foi extraída da organização "Our World in Data", que consolida estatísticas globais sobre a COVID-19.
    - **URL:**https://ourworldindata.org/grapher/cumulative-deaths-and-cases-covid-19?overlay=download-data
3. **Dados de População Total:** Os dados demográficos anuais, contidos no ficheiro *population_totals.csv*, foram recolhidos a partir do indicador "Population, total" do World Bank.
    - **URL:** https://data.worldbank.org/indicator/SP.POP.TOTL