**Sprint – Engenharia de Dados**

**Relatório: MVP** - **Análise da COVID-19**

**Aluna: Fernanda dos Santos Soares**

1. **Introdução**

Este MVP visa criar um **pipeline de dados na nuvem utilizando o Databricks**, com estrutura em modelo estrela, que integre informações sobre **casos, mortes e vacinação** por país e data, permitindo análises comparativas. Os dados foram coletados do Kaggle e carregados diretamente no Databricks Community Edition.

1. **Objetivo Geral**

Desenvolver um MVP de engenharia de dados na nuvem, utilizando o Databricks, para consolidar e analisar dados sobre a pandemia de COVID-19, com foco em compreender os impactos da vacinação na evolução de casos e mortes, através de um modelo de dados analítico em estrela  
  
**2.1 Objetivos Específicos**

1. Realizar a ingestão, tratamento e transformação dos dados de casos/mortes e vacinação no Databricks.
2. Modelar os dados utilizando a abordagem dimensional (modelo estrela), com fato e dimensões adequadas.
3. Integrar as duas bases de dados em um modelo analítico único, utilizando chaves comuns como país e data.
4. Construir visualizações e análises exploratórias para responder perguntas relevantes sobre a pandemia.
5. Identificar padrões e possíveis correlações entre o avanço da vacinação e a redução de casos/mortes.
6. Avaliar a qualidade e integridade dos dados utilizados.
   1. **Objetivos das análises**

As análises realizadas tiveram como propósito:

1. Identificar os países mais impactados em termos de número de casos e mortes;
2. Mapear os países com maior esforço de vacinação, tanto em números absolutos quanto proporcionais à população;
3. Detectar correlações entre vacinação e mortalidade, avaliando a efetividade da vacinação no controle da doença;
4. Apontar exceções e padrões fora da média, como países com alta mortalidade mesmo com vacinação elevada, ou países com baixa vacinação e alto impacto;
5. Analisar a distribuição de fabricantes de vacinas e identificar quais são os mais utilizados nos países com maior taxa de vacinação.

Com base nessas análises, o trabalho oferece evidências para compreender o papel da vacinação no enfrentamento da pandemia, além de destacar vulnerabilidades e boas práticas globais que podem orientar futuras decisões em saúde pública.

1. **Tecnologias Utilizadas**

* **Databricks Community Edition** – Plataforma de desenvolvimento e execução do pipeline de dados.
* **Apache Spark / PySpark** – Processamento distribuído e manipulação de dados com DataFrames.
* **Delta Lake** – Armazenamento em camadas (Bronze, Silver e Gold), garantindo versionamento e integridade dos dados.
* **SQL (no Databricks**) – Consultas para transformação, análise e validação dos dados.
* **Conjuntos de dados Kaggle** – Bases de dados públicas sobre COVID-19 e vacinação

[COVID-19 World Vaccination Progress](https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress)

[Covid Cases and Deaths WorldWide](https://www.kaggle.com/datasets/themrityunjaypathak/covid-cases-and-deaths-worldwide?resource=download)

Licença de uso open.

* **Modelagem Estrela** – Estrutura de dados organizada em fato e dimensões para facilitar análises.
* **Visualizações com PySpark + Pandas + Matplotlib** – Criação de gráficos para interpretação dos dados.

1. **Modelagem e Catálogo de dados**

Foi adotado um modelo em estrela, com tabelas factuais consolidadas na camada Gold. As principais tabelas construídas foram:

* vacinas\_total\_por\_pais\_fabricante
* casos\_mortes\_por\_pais
* vacinas\_casos\_mortes\_flat

Abaixo segue o Catálogo de Dados contendo os principais atributos das tabelas utilizadas.

***Tabela: vacinas\_total\_por\_pais\_fabricante***

* pais (string): Nome do país
* fabricante (string): Nome do fabricante da vacina
* total\_vacinas (long): Total de vacinas aplicadas por fabricante no país

***Tabela: casos\_mortes\_por\_pais***

* pais (string): Nome do país
* total\_casos (long): Total de casos registrados
* total\_mortes (long): Total de mortes registradas
* population (long): População estimada do país

***Tabela: vacinas\_casos\_mortes\_flat***

* pais (string): Nome do país
* fabricante (string): Nome do fabricante da vacina
* total\_vacinas (long): Total de vacinas aplicadas
* total\_casos (long): Total de casos registrados
* total\_mortes (long): Total de mortes registradas
* population (long): População estimada do país

1. **Organização do repositório**

* notebook/: notebook completo com o pipeline e análises
* dataset/: referência da fonte de dados usada (Kaggle)

1. **Análise e Evidências**

A seguir, foram conduzidas análises com base nos dados consolidados da base **gold**, que contempla informações globais sobre casos, mortes e vacinação contra a COVID-19. As análises permitiram levantar evidências relevantes a partir de perguntas-chave, descritas abaixo:

1. **Quais são os países com maior número de casos e mortes de COVID-19?**  
   A análise dos dados mostrou que os 10 países com mais casos são: **Estados Unidos, Índia, França, Alemanha, Brasil, Japão, Coreia do Sul, Itália, Reino Unido e Rússia**. Os Estados Unidos lideram tanto em número de casos quanto em mortes. A alta incidência nesses países pode ser associada à densidade populacional, políticas de controle e estágios da pandemia.

Gráfico

Descrição gerada automaticamente

1. **Quais países aplicaram o maior número de vacinas?**

Os países que mais aplicaram vacinas são **Alemanha, França, Italia Argentina, Peru, Chile, entre outros.**Gráfico, Gráfico de barras

Descrição gerada automaticamente

1. **Qual a taxa de vacinação por habitante em cada país?**

Gráfico, Histograma

Descrição gerada automaticamente

1. **Quais países têm alta mortalidade e baixa vacinação (fora da média)?**  
   Gráfico

   Descrição gerada automaticamente

Tabela

Descrição gerada automaticamente

1. **Qual fabricante de vacina é o mais utilizado em cada país?**  
   A análise do fabricante com maior volume de aplicação por país revelou padrões regionais interessantes. Por exemplo, **Pfizer-BioNTech** lidera em países da Europa e América do Norte, enquanto **Sinopharm** e **Sinovac** são predominantes em nações da Ásia e África. No Brasil, a vacina mais aplicada foi a **AstraZeneca**.

**Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente**

1. **Autoavaliação**Durante a execução do projeto, o objetivo principal foi construir um pipeline de engenharia de dados funcional, aplicar conceitos de transformação com modelo estrela e criar análises de valor com base em dados reais. Com base no trabalho concluído, acredito que os objetivos foram atingidos.

Entre as principais **dificuldades enfrentadas**, destacam-se:

* Tratamento de dados inconsistentes: as bases exigiram múltiplas limpezas e conversões, especialmente para garantir integridade nas colunas numéricas.
* Problemas com schemas e partições: alguns erros no Databricks, como conflitos de schema, exigiram ajustes de configuração e uso de mergeSchema.
* União de tabelas complexas: por envolver junções de fontes distintas, foi necessário alinhar nome de colunas e formatos para garantir que as queries funcionassem.

Apesar das dificuldades, o aprendizado técnico foi enriquecedor, especialmente no uso do PySpark e do Databricks como ferramenta de processamento em larga escala.

1. **Trabalhos Futuros**

Como sugestões para evolução do projeto:

* Adicionar a dimensão temporal ao modelo, para análise de tendências mais apuradas.
* Explorar visualizações mais avançadas, com dashboards interativos em Power BI ou Plotly.
* Incorporar novas variáveis, como lockdowns, taxa de testagem ou mobilidade populacional.
* Utilizar machine learning para prever evolução de casos com base no histórico por país.