**Decifrando a COVID-19: Insights de Business Intelligence em Dados Pandêmicos e Políticas Públicas**

**1 INTRODUÇÃO**

**1.1 Apresentação do tema**

A pandemia do COVID-19 é um evento global sem precedentes na era moderna, impactando profundamente a saúde, a economia e a vida social em todo o mundo. Desde seu surgimento, houve uma corrida para entender a doença, mitigar sua propagação e desenvolver estratégias de resposta. Neste cenário, o Business Intelligence (BI) emergiu como uma ferramenta crítica, fornecendo insights específicos para governos e organizações na tomada de decisões baseadas em dados.

A aplicação de BI em saúde pública pode ajudar a melhorar a tomada de decisão, a eficiência dos serviços e a qualidade dos cuidados prestados. No contexto da pandemia da COVID-19, o BI pode ser utilizado para rastrear a evolução da doença, avaliar a eficácia das políticas públicas e identificar grupos de risco.

“Do ponto de vista assistencial e humanitário – no combate direto à proliferação do vírus e no planejamento operacional de governos e instituições de saúde –, e também como arma das empresas para passar pela crise econômica que se projeta como resultado da pandemia de Covid-19, o uso de ferramentas de análise de dados é fundamental para antecipar e projetar cenários que possam embasar as melhores decisões para contornar os impactos econômicos e sociais.” (Maia, 2020)

Em síntese, a pandemia da COVID-19 revelou a imperatividade de adotar uma abordagem fundamentada em dados não apenas como uma resposta assistencial e humanitária, mas também como uma ferramenta estratégica essencial para enfrentar os desafios econômicos. O emprego essencial de ferramentas de análise de dados, como analytics, big data e inteligência artificial, destaca-se como crucial na capacidade de antecipar cenários, sustentar decisões eficazes e mitigar os impactos econômicos e sociais resultantes desta crise global.

Neste contexto, a análise de dados transcende seu papel informativo, tornando-se uma questão de sobrevivência para setores assistenciais, econômicos e humanitários. A capacidade de resolver e antecipar problemas emerge como essencial para a adaptação e superação da crise, reforçando a relevância fundamental da análise de dados como alicerce para uma resposta eficiente e resiliente em períodos críticos como o enfrentado atualmente.

**1.2 Justificativa da escolha do tema**

Em um cenário de crise global como a pandemia da COVID-19, o emprego estratégico de ferramentas de análise de dados revela-se imprescindível. Do ponto de vista humanitário, a capacidade de identificar movimentações e aglomerações é vital para conter a propagação do vírus, enquanto a gestão eficaz de recursos médicos e hospitalares torna-se possível com a aplicação de Business Intelligence (BI). A análise de dados também se estende à esfera epidemiológica, auxiliando na identificação de padrões e previsões cruciais para o desenvolvimento de vacinas e medicamentos.

Dessa forma, este tema foi escolhido para demonstrar a aplicabilidade do BI não apenas no contexto comercial tradicional, mas também na esfera da saúde pública, onde decisões informadas podem salvar vidas. O uso de ferramentas avançadas como Apache Hop, Jenkins, SQL Server e Power BI para extrair, transformar, carregar (ETL) e visualizar dados destaca a multifuncionalidade do BI e sua capacidade de lidar com grandes volumes de dados complexos - um desafio inerente à pandemia.

**2 OBJETIVOS**

**2.1 Objetivo geral**

O objetivo geral deste estudo é explorar o papel do Business Intelligence (BI) na análise de dados da COVID-19 para extrair insights significativos que possam informar e melhorar as políticas públicas em resposta à pandemia. Através de um estudo de caso detalhado que utiliza dados públicos, este trabalho visa demonstrar como as técnicas de BI ajudam na compreensão e na gestão de uma crise de saúde global.

**2.2 Objetivos Específicos**

1. Analisar a Progressão da Pandemia: Utilizar as ferramentas de BI para analisar os dados de progressão da pandemia, tais como os casos confirmados, as mortes e as taxas de recuperação, identificando padrões e tendências que possam indicar a eficácia das medidas tomadas (Grupo de Modelos Matemáticos para COVID-19 do Vale do São Francisco, 2021) .
2. Avaliar Políticas Públicas: Comparar a temporalidade e severidade das políticas públicas, como lockdowns e campanhas de vacinação, com as mudanças nos dados de casos e mortalidade da COVID-19, determinando a relação entre ação governamental e resultados de saúde pública (Autor Desconhecido, 2021).
3. Estimar a Distribuição de Recursos: Investigar como o apoio financeiro foi distribuído durante a pandemia e se há uma revelação entre a ajuda financeira e a mitigação dos efeitos da COVID-19 nas comunidades afetadas, usando as bases de dados fornecidas.
4. Aprimorar a Visualização de Dados Pandêmicos: Desenvolver dashboards interativos e relatórios no Power BI que podem ser usados ​​pelos responsáveis ​​pela tomada de decisão para entender melhor a situação atual e histórica da pandemia, permitindo respostas mais ágeis e informadas.
5. Contribuir para a Educação em Saúde Pública: Fornecer um recurso educacional que possa ser utilizado para instruir profissionais de saúde pública e a população em geral sobre a importância da análise de dados durante crises de saúde, realçando o impacto do BI na compreensão e resposta a pandemias.
6. Fomentar a Tomada de Decisão Baseada em Dados: Avaliar como as técnicas de BI podem ser incorporadas nas estratégias de tomada de decisão dos órgãos de saúde pública para melhorar a resposta a crises sanitárias futuras.

**2 REFERENCIAL TEÓRICO**

**2.1 Business Intelligence**

\*Conceito e descrição do BI aqui!

**2.2 Data Warehouse**

Um Data Warehouse (DW) é um sistema de armazenamento de dados centralizado e integrado que é projetado para dar suporte a análises e tomada de decisões. Ele é construído para consolidar dados de várias fontes, transformá-los e armazená-los em um formato adequado para análise e relatórios.

O objetivo do DW é fornecer acesso a informações precisas e confiáveis ​​para suportar a tomada de decisões estratégicas e táticas de negócios. Ele é projetado para responder a consultas complexas que não podem ser facilmente respondidas pelos sistemas transacionais de uma empresa, como sistemas de gestão empresarial (ERP) ou sistemas de gerenciamento de relacionamento com o cliente (CRM).

Os DWs geralmente incluem uma série de etapas de processamento de dados, como extração, transformação e carga (ETL). Isso envolve a extração de dados de várias fontes, como bancos de dados operacionais, aplicativos de terceiros e fontes externas, a transformação dos dados em um formato padronizado e a carga dos dados em um repositório centralizado.

Os dados são geralmente organizados em torno de temas de negócios, como vendas, marketing, finanças ou recursos humanos, para facilitar a análise e a geração de relatórios. Os dados são armazenados em um formato otimizado para consultas e análises, geralmente em um modelo dimensional, que é projetado para suportar as necessidades de relatórios e análises de negócios.

Os DWs também são projetados para suportar análises de negócios mais avançadas, como mineração de dados, análise preditiva e análise de séries temporais. Essas técnicas permitem que as empresas identifiquem tendências e padrões em grandes conjuntos de dados e usem essas informações para tomar decisões informadas sobre o futuro do negócio.

Em resumo, um DW é um sistema centralizado e integrado de armazenamento de dados que é projetado para dar suporte a análises e tomada de decisões de negócios. Ele é construído para consolidar dados de várias fontes, transformá-los em um formato adequado para análise e relatórios e suportar consultas complexas e análises avançadas.

**2.3 Data Mart**

Data Mart é uma abordagem de armazenamento de dados que envolve a criação de um repositório de dados específico para um departamento, processo ou aplicação específica. Ao contrário de um Data Warehouse (DW), que armazena dados de toda a organização, um Data Mart é focado em um subconjunto específico de dados.

O objetivo do Data Mart é fornecer acesso a informações específicas e relevantes para uma área específica de negócios, permitindo que os usuários finais realizem análises e tomem decisões informadas. Isso é feito através da integração de dados de várias fontes e da organização dos dados em torno de temas de negócios específicos.

Os Data Marts podem ser criados a partir de dados que já estão armazenados em um DW ou podem ser criados independentemente, dependendo das necessidades de negócios. Eles geralmente envolvem a extração, transformação e carregamento (ETL) de dados de fontes operacionais, bem como a agregação de dados em um formato adequado para análises específicas.

Os Data Marts podem ser projetados para suportar consultas ad hoc, relatórios pré-definidos e análises avançadas, como mineração de dados e análise preditiva. Eles também podem ser projetados para suportar o acesso direto do usuário final, permitindo que os usuários finais criem suas próprias consultas e relatórios.

Em resumo, um Data Mart é um repositório de dados específico para um departamento, processo ou aplicação específica. Eles são projetados para fornecer acesso a informações específicas e relevantes para uma área específica de negócios, permitindo que os usuários finais realizem análises e tomem decisões informadas. Eles podem ser criados a partir de dados armazenados em um DW ou independentemente e podem ser projetados para suportar consultas ad hoc, relatórios pré-definidos e análises avançadas.

**2.4 ETL**

**\***Conceito e definição de ETL aqui!

**2.5 OLAP**

Cubo OLAP (On-line Analytical Processing) é uma técnica de processamento de dados multidimensional usada para a análise de dados em Business Intelligence (BI). É uma estrutura de dados que permite que os dados sejam visualizados a partir de diferentes perspectivas, incluindo dimensões, hierarquias e medidas.

Os cubos OLAP são compostos por fatos, que são os valores numéricos que se deseja analisar (por exemplo, vendas) e dimensões, que são as categorias ou atributos pelos quais os fatos são analisados (por exemplo, tempo, produto ou região). Os cubos OLAP também podem incluir hierarquias, que são as estruturas de níveis que definem a relação entre as dimensões (por exemplo, ano, trimestre, mês) e medidas, que são as fórmulas aplicadas aos fatos para calcular dados resumidos (por exemplo, somas, médias e desvios padrão).

Os cubos OLAP podem ser armazenados em Data Warehouses ou em bases de dados multidimensionais especializadas, projetadas para suportar consultas OLAP. Eles permitem que os usuários explorem os dados de uma forma intuitiva, arrastando e soltando dimensões e medidas em diferentes eixos de um cubo. As consultas são resolvidas rapidamente, pois os dados são pré-agregados e organizados em uma estrutura de dados otimizada para consulta multidimensional.

Os cubos OLAP são uma importante ferramenta de análise de dados em Business Intelligence, permitindo que os usuários analisem grandes volumes de dados de uma forma rápida e fácil, obtendo insights sobre o desempenho do negócio e identificando tendências e oportunidades de melhoria. Eles são amplamente utilizados em diversos setores, incluindo varejo, finanças, manufatura e saúde, entre outros.

**2.6 Big Data**

**\***Conceito e definição de Big Data aqui!

**3 FERRAMENTAS UTILIZADAS**

A escolha das ferramentas e metodologias para a análise de dados complexos não é uma decisão que deve ser tomada levianamente, especialmente no contexto de uma crise de saúde pública global como a pandemia de COVID-19. Para este estudo, a seleção de ferramentas é orientada pela necessidade de executar processos de ETL robustos, armazenamento de dados eficaz, integração contínua e visualização de dados interativos e informativos.

**3.1 Apache Hop - Hop Orchestration Platform**

O **Apache Hop** (Hop Orchestration Platform) foi escolhido pela sua capacidade de facilitar a orquestração de fluxos de dados complexos, permitindo uma integração eficiente de diversas fontes de dados. Sua natureza leve e flexível é ideal para o processamento de grandes conjuntos de dados do COVID-19, proporcionando uma plataforma agnóstica para a transformação de dados que pode se adaptar rapidamente às mudanças nas necessidades de análise.

De acordo com a documentação do Apache Hop, ele é:

“(...) uma plataforma de orquestração e engenharia de dados que visa facilitar todos os aspectos da orquestração de dados e metadados. Hop permite que você se concentre no problema que está tentando resolver sem que a tecnologia atrapalhe. Tarefas simples devem ser fáceis, tarefas complexas precisam ser possíveis.” (Equipe Hop, 2019-2023)

**3.2 SQL Server**

O **SQL Server** foi selecionado como o sistema de gerenciamento de banco de dados (SGBD) devido à sua robustez, segurança e escalabilidade, que são essenciais para gerenciar o volume e a sensibilidade dos dados relacionados à saúde. Sua compatibilidade com diversas ferramentas de BI e capacidade de realizar consultas complexas e análises de dados tornam-no uma escolha confiável para armazenar e gerenciar os dados que servirão de base para este estudo.

De acordo com a documentação do SQL Server, ele é: **#TODO**

**3.3 Jenkins**

O **Jenkins** é uma ferramenta de automação de código aberto que oferece suporte à melhoria de integração contínua (CI) e entrega contínua (CD). A integração do Jenkins no fluxo de trabalho permite a atualização e o teste automatizado dos dados, o que é crucial para garantir a precisão e a atualidade das informações em um cenário em rápida evolução como a pandemia.

De acordo com a documentação do Jenkins, ele é: **#TODO**

**3.4 Microsoft Power BI**

O **Power BI** é uma ferramenta de visualização de dados intuitiva e poderosa que oferece a capacidade de criar dashboards interativos e relatórios abrangentes. A escolha desta ferramenta é justificada pela sua facilidade de uso, como opções avançadas de especificidade e a capacidade de compartilhar insights de forma eficaz com partes interessadas, que podem variar de decisões políticas para profissionais de saúde.

“O Power BI tem importantes características que facilitam a manipulação de grandes volumes de dados, pois, possibilita importação de diferentes fontes de dados permitindo o relacionamento entre elas. As informações brutas que o usuário utilizará no software, relatórios que organizam os dados em um conjunto de gráficos e painéis atualizados, que fornecem visualizações específicas com base nesses relatórios, podendo ainda ser publicados e compartilhados com vários usuários através da internet, Desktops e até mesmo adaptados para smartphones, proporcionam cada vez mais a praticidade que o mundo corporativo necessita atualmente” (FRAGA, 2019, p. 19)

**4. METODOLOGIA**

Pode-se dizer que uma pesquisa é realizada quando há a intenção de encontrar soluções para questões que não possuem respostas imediatas. Conforme Gil (2002, p. 17), “Pode-se definir pesquisa como o procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos. A pesquisa é requerida quando não se dispõe de informação suficiente para responder ao problema, ou então quando a informação disponível se encontra em tal estado de desordem que não possa ser adequadamente relacionada ao problema.”

Seguindo a classificação do autor em relação aos tipos de pesquisa, é possível concluir que este trabalho é caracterizado como uma pesquisa exploratória-descritiva. Essa abordagem proporciona uma maior familiaridade com o tema proposto, apresentando diferentes perspectivas de diversos autores e descrevendo com precisão os dados pandêmicos, utilizando instrumentos padronizados de coleta de dados, que conduzem a resultados de natureza quantitativa. (GIL, 2002).

A metodologia a ser empregada é baseada em princípios de Business Intelligence e análise de dados, extraindo o máximo de valor dos dados disponíveis. A abordagem aplicável permitirá o tratamento, a análise e a interpretação dos dados de maneira a responder às perguntas de pesquisa de forma objetiva e fundamentada.

Esta pesquisa compreendeu uma revisão bibliográfica sobre o tema, utilizando websites, artigos, livros e dissertações como fontes. A natureza recente do tema abordado, que é a pandemia da COVID-19, resulta em uma escassez de produções acadêmicas e/ou livros sobre o assunto. Adicionalmente, foram conduzidas análises nos websites das Secretarias de Saúde de todos os estados do país para verificar e coletar dados sobre o período pandêmico.

**4.1 Descrição do processo de extração de dados do site Our World in Data**

O processo de extração de dados para este estudo envolveu a aquisição sistemática de conjuntos de dados relacionados à pandemia de COVID-19 do repositório globalmente reconhecido Our World in Data. Este repositório é conhecido por compilar e disponibilizar dados de alta qualidade, que são encontrados de fontes confiáveis ​​e atualizadas regularmente, fornecendo uma base sólida para análises epidemiológicas e socioeconômicas.

A seleção dos dados foi guiada pela relevância para as questões de pesquisa do estudo, focando em dois domínios principais: dados epidemiológicos e registros de políticas públicas rompidas. Os conjuntos de dados específicos baixados em formato CSV incluem:

Dados Epidemiológicos e de Saúde:

**#TODO: adicionar links das páginas onde os CSVs foram baixados.**

2-casos-covid-data: Dados diários de casos confirmados de COVID-19.

3-explorador-data: Métricas gerais relacionadas à pandemia, incluindo testes e resultados de casos.

5-hospitalizacoes\_por\_semana-covid: Informações sobre hospitalizações semanais devido à COVID-19.

6-mortes-covid-data: Números diários e acumulados de mortes atribuídas à COVID-19.

9-risco-mortalidade-covid-data: Estatísticas sobre o risco de mortalidade e fatores de risco associados.

10-testes-covid-data: Dados sobre a quantidade e resultado dos testes de COVID-19 realizados.

12-vacinacao-covid-data: Dados sobre a distribuição e administração de vacinas contra a COVID-19.

Registros de Políticas Públicas:

**#TODO: adicionar links das páginas onde os CSVs foram baixados.**

1-apoio-financeiro: Informações sobre apoio financeiro e medidas de estímulo econômico.

4-Fechamento\_Escolas\_E\_Locais\_de\_trabalho-covid: Dados sobre fechamento de escolas e locais de trabalho como medidas de contenção.

7-politicas-de-vacinacao: Detalhes sobre políticas de vacinação e diretrizes inovadoras.

8-restricoes-ficar-em-casa-covid: Dados sobre ordens de permanência em casa e confinamentos.

11-uso-de-mascaras-policies-covid: Registros de políticas relacionadas ao uso de máscaras e outras proteções visuais.

A extração dos dados foi realizada manualmente, seguindo estes passos:

1. Navegação até o site Our World in Data e localização da seção dedicada aos dados da COVID-19.
2. Identificação e seleção dos conjuntos de dados relevantes para o estudo.
3. Download dos arquivos CSV correspondentes aos conjuntos de dados selecionados.
4. Verificação dos arquivos baixados para garantir a integridade dos dados.
5. Validação e Preparação de Dados.
6. Remoção, onde os dados foram submetidos a um processo de validação e preparação, incluindo: Verificação da consistência das séries temporais e dos marcadores de identificação geográfica; Limpeza de possíveis discrepâncias ou entradas ausentes; Padronização dos formatos de dados e outros campos relevantes para análise.

Esta metodologia de extração de dados garante que uma análise subsequente seja realizada em uma base de dados sólida e confiável, permitindo a realização de uma investigação metodológica rigorosa e a obtenção de resultados significativos.

**4.2 Utilização do Apache Hop para orquestração do fluxo de dados**

A orquestração do fluxo de dados é uma etapa crítica em qualquer processo de Business Intelligence, especialmente quando lidamos com grandes volumes de dados e a necessidade de integrar várias fontes de maneira eficiente e confiável. Para atender a essa necessidade, o Apache Hop (Hop Orchestration Platform) foi escolhido devido à sua eficiência, flexibilidade e abordagem orientada por metadados na orquestração de pipelines de dados.

O Apache Hop foi configurado para automatizar o fluxo de trabalho de ETL (Extração, Transformação e Carga), seguindo estes passos principais:

1. Configuração do Ambiente:

* Instalação do Apache Hop e configuração do ambiente de desenvolvimento em uma máquina local ou servidor dedicado.
* Definição de variáveis de ambiente e conexões de banco de dados para o SQL Server, onde os dados extraídos serão armazenados.

1. Desenvolvimento de Pipelines:

* Criação de projetos e ambientes dentro do Hop para separar e gerenciar os fluxos de trabalho de acordo com os conjuntos de dados específicos.
* Desenvolvimento de pipelines dedicados à redução de cada conjunto de dados CSV, utilizando os recursos do Hop para leitura, transformação e escrita de dados.
* Implementação de transformações para limpeza e normalização dos dados, garantindo a qualidade e consistência para análises subsequentes.

1. Orquestração e Agendamento:

* Utilização do Apache Hop para orquestrar a execução sequencial ou paralela dos pipelines, garantindo que os dados sejam processados na ordem correta.
* Agendamento de pipelines para execução automatizada, permitindo a atualização regular dos dados no banco de dados SQL Server, o que é vital para manter os dashboards e relatórios atualizados.

1. Monitoramento e registro:

* Configuração de monitoramento dentro do Hop para rastrear o desempenho dos pipelines e identificar quaisquer falhas ou erros durante a execução.
* Implementação de logs detalhados para cada pipeline, facilitando a depuração e a auditoria dos processos de ETL.

1. Integração Contínua:

* Integração do Apache Hop com ferramentas de integração contínua, como o Jenkins, para automatizar a implantação e a execução de pipelines em ambientes de produção.

O uso do Apache Hop para orquestrar o fluxo de dados fornece uma base sólida e escalável para o processamento de dados relacionados ao COVID-19. A capacidade de adaptar rapidamente os dutos em resposta a novas informações e mudanças nos dados é uma vantagem significativa no contexto sonoro de uma pandemia. A metodologia adotada com o Apache Hop garante que os dados sejam coletados e preparados de forma que permitam a realização de uma análise confiável e abrangente nas etapas subsequentes do projeto.

**4.3 Configuração do Jenkins para integração contínua dos dados**

A integração contínua (CI) é um pilar central na estratégia de Business Intelligence para garantir que os dados utilizados nas análises sejam sempre atuais e confiáveis. Jenkins, uma ferramenta de automação de código aberto, é empregada para facilitar o CI, automatizando a atualização dos dados processados ​​pelo Apache Hop no banco de dados SQL Server.

Etapas de Configuração no Jenkins:

1. Instalação do Jenkins:

* Instalação e configuração do Jenkins em um servidor confiável e seguro.
* Configuração dos plugins necessários para integrar o Jenkins com o Apache Hop e o SQL Server.

1. Criação de empregos de CI:

* Configuração de um 'job' de CI no Jenkins para cada pipeline de ETL desenvolvido no Apache Hop.
* de gatilhos de execução com base em cronograma ou detecção de mudanças nos dados de entrada, garantindo que os dados sejam atualizados após qualquer nova carga de dados ou em intervalos regulares.

1. Automatização do Processo de ETL:

* Configuração de 'jobs' para executar automaticamente os processos de ETL, transformando e carregando os dados no SQL Server.
* Implementação de procedimentos para validar a integridade dos dados após cada atualização, garantindo a acurácia dos dados no banco de dados.

1. Monitoramento e Notificações:

* Configuração de notificações para informar a equipe responsável sobre o sucesso ou falha nos processos de ETL.
* Uso de dashboards do Jenkins para monitorar a saúde do processo de CI e fornecer visibilidade sobre o pipeline de dados.

1. Manutenção e Ajustes:

* Estabelecimento de rotinas de manutenção para atualizar e ajustar os 'trabalhos' de CI conforme necessário, garantindo que eles sejam alinhados com as mudanças nos requisitos de dados ou na lógica de ETL.
* Monitoramento contínuo do desempenho e da eficácia dos 'trabalhos' de CI para otimização contínua.

A implementação do CI com Jenkins é fundamental para o sucesso do projeto, garantindo que os dados mais recentes e precisos estejam sempre disponíveis para análise e tomada de decisão. Isso é especialmente crítico no contexto dinâmico da pandemia de COVID-19, onde a situação epidemiológica e as respostas políticas evoluem rapidamente. A configuração do Jenkins permite a adaptação e resposta rápida a essas mudanças, garantindo que os dashboards e relatórios reflitam a situação mais atualizada possível.

**4.4 Armazenamento e gerenciamento de dados usando o SQL Server**

Após a orquestração e execução automatizada dos processos de ETL via Apache Hop e Jenkins, o próximo passo crucial é o armazenamento e gerenciamento eficiente dos dados. Para este fim, o SQL Server, um sistema de gerenciamento de banco de dados relacionado da Microsoft, foi escolhido por sua robustez, segurança e conformidade com as necessidades de uma análise de dados complexa e escalável.

Etapas de armazenamento e gerenciamento dos dados:

1. Configuração do Banco de Dados:

* Criação de um banco de dados dedicado no SQL Server para hospedar os conjuntos de dados do COVID-19.
* Estruturação de tabelas normalizadas para refletir as diversas dimensões dos dados, como casos, mortalidade, políticas de saúde pública e dados de vacinação.
* Definição de índices e chaves estrangeiras para otimizar consultas e manter a integridade referencial dos dados.

1. Processo de Carga de Dados:

* Utilização de procedimentos armazenados e tarefas de ETL programadas no Jenkins para carregar dados no banco de dados de forma automática e recorrente.
* Implementação de transações para garantir que as transações de carga de dados sejam atômicas, consistentes, isoladas e seguras (ACID).

1. Segurança e Conformidade:

* Configuração de políticas de segurança, incluindo autenticação e autorização, para garantir que somente usuários autorizados possam acessar ou modificar os dados.
* Ativação de recursos de auditoria e conformidade para monitorar o acesso aos dados e garantir o cumprimento das regulamentações de privacidade e proteção de dados.

1. Backup e Recuperação:

* Estabelecimento de uma estratégia de backup para proteger os dados contra perda ou corrupção, incluindo backups completos e incrementais.
* Planejamento de uma estratégia de recuperação de desastres para garantir a continuidade do acesso aos dados em caso de falhas de sistema ou outros incidentes.

1. Otimização e Manutenção:

* Monitoramento regular do desempenho do banco de dados e ajuste de consultas e índices para garantir tempos de resposta rápidos e eficientes.
* Programação de rotinas de manutenção, como a reindexação e a atualização de estatísticas, para manter o desempenho ideal do banco de dados ao longo do tempo.

**4.5 Desenvolvimento de dashboards e relatórios no Power BI**

Com os dados processados, orquestrados e armazenados de maneira eficiente, o próximo passo metodológico é a construção de dashboards e relatórios no Power BI. Esta etapa é crucial para a visualização de dados, o que permite uma compreensão rápida e intuitiva das complexidades e tendências inerentes aos dados da COVID-19.

1. Estratégia de Visualização:

* Identificação das principais informações e KPIs (Key Performance Indicators) que serão o foco dos dashboards, como número de casos, taxas de vacinação e eficácia das políticas de saúde pública.
* Definição de uma estratégia de visualização que atende às necessidades dos usuários finais, sejam eles pesquisadores, profissionais de saúde pública ou decisões políticas.

1. Desenvolvimento dos Dashboards:

* Criação de um modelo de dados no Power BI que reflita a estrutura e as relações dos dados armazenados no SQL Server.
* Utilização de funcionalidades avançadas do Power BI para desenvolver dashboards interativos e relatórios que facilitam a análise de dados, como slicers, drill-downs e gráficos de tempo real.
* Design que segue as melhores práticas de usabilidade e visualização de dados, garantindo que os dashboards sejam acessíveis e compreensíveis para todos os públicos.

1. Teste e Iteração:

* Realização de testes rigorosos com usuários para garantir a precisão dos dashboards e a clareza das visualizações.
* Coleta de feedback para iterar e aprimorar os dashboards, garantindo que eles atendam às necessidades de evolução das partes interessadas.

1. Implementação e Compartilhamento:

* Publicação dos dashboards no serviço Power BI online para fácil acesso e compartilhamento com usuários autorizados.
* Configuração de gateways de dados para garantir que os dashboards e relatórios sejam atualizados automaticamente com os dados mais recentes.

1. Manutenção e Evolução:

* Monitoramento contínuo e manutenção dos dashboards para garantir que continuem a fornecer insights relevantes à medida que novos dados se tornem disponíveis.
* Adaptação e evolução dos dashboards para incorporar novas fontes de dados ou para refletir mudanças nas políticas de saúde pública e na evolução da pandemia.

O desenvolvimento de dashboards e relatórios no Power BI é um componente essencial do projeto, transformando dados brutos em visualizações que informam e capacitam os envolvidos na resposta à pandemia de COVID-19. Esta etapa finaliza o ciclo de BI, onde os dados não apenas informam sobre o passado e o presente, mas também fornecem a capacidade de prever e preparar-se para futuras tendências e surtos.

**5. Execução do Projeto (incorporar na metodologia, talvez?)**

**5.1 Detalhamento das etapas de implementação das ferramentas**

Após as instalações e configurações preliminares, a implementação das ferramentas de BI no projeto avança para uma abordagem sistemática e iterativa, garantindo a integração e funcionalidade efetivas para análise dos dados da COVID-19.

5.1.1 Desenvolvimento de Pipelines com Apache Hop:

Estruturação dos Pipelines: Construção de pipelines que refletem o fluxo de dados desde a remoção dos CSVs até a carga no SQL Server, com ênfase na modularidade e reutilização de componentes.

Transformação e Enriquecimento de Dados: Aplicação de transformações lógicas para enriquecer os dados com informações adicionais, como cálculos de tendências e taxas de crescimento, que são essenciais para a análise subsequente.

5.1.2 Automatização de Fluxos de Trabalho com Jenkins:

Orquestração de Pipelines: Implementação da orquestração de pipelines, utilizando 'jobs' de Jenkins para automatizar as execuções, garantindo uma atualização constante dos dados no sistema.

Monitoramento de Execução: Configuração de mecanismos de monitoramento e alertas para garantir que quaisquer falhas nos pipelines sejam rapidamente identificadas e resolvidas, mantendo a integridade do fluxo de dados.

5.1.3 Gerenciamento de Dados no SQL Server:

Integridade de Dados: Implementação de procedimentos armazenados e funções definidas pelo usuário para manter a integridade e consistência dos dados, aplicando regras de negócio e validações.

Desempenho e Otimização: Monitoramento do desempenho do banco de dados e ajustes nas consultas e estruturas de dados para melhorar os tempos de resposta e a eficiência do armazenamento de dados.

5.1.4 Visualização com Power BI:

Construção de Dashboards: Desenvolvimento de dashboards que sintetizam os dados processados, utilizando visualizações como mapas de calor, gráficos de linha e barra, e cartões de indicadores para destacar informações críticas.

Interatividade e Usabilidade: Incorporação de recursos interativos que permitem aos usuários filtrar e examinar os dados em diferentes níveis, desde visões globais até análises granulares por região ou período.

**5.2 Processo de integração e limpeza dos dados**

A integração e a limpeza dos dados são etapas fundamentais no projeto de Business Intelligence, especialmente quando tratamos de grandes conjuntos de dados oriundos de múltiplas fontes, como é o caso dos dados da pandemia de COVID-19. Este processo é vital para garantir a precisão das análises e a confiabilidade das conclusões extraídas dos dados.

5.2.1 Etapas de Integração dos Dados:

Consolidação: Após a remoção, os dados de diversas fontes foram consolidados em um único repositório, utilizando o SQL Server como o sistema centralizado para armazenamento.

Padronização: Os dados foram padronizados para garantir a uniformidade dos formatos de dados, nomenclatura de variáveis ​​e unidades de medida, facilitando comparações e análises cruzadas.

Verificação de Consistência: A consistência entre conjuntos de dados foi verificada, incluindo a correspondência de códigos de países e regiões, para garantir que todos os dados possam ser integrados corretamente.

5.2.2 Etapas de Limpeza dos Dados:

Remoção de Duplicatas: As entradas duplicadas foram identificadas e removidas, evitando distorções nas análises.

Tratamento de Valores Ausentes: Foram aplicadas técnicas para tratar valores ausentes, como a imputação com base em médias ou medianas, ou a exclusão de registros quando a imputação não era adequada.

Detecção e Correção de Anomalias: Foram utilizados métodos estatísticos e visualizações para detectar anomalias nos dados, como valores que fogem dos padrões esperados, e correções foram aplicadas quando necessário.

5.2.3 Validação dos Dados:

Confirmação de Fontes: As fontes dos dados foram revisadas para confirmar sua veracidade e confiabilidade.

Verificação de Qualidade de Dados: Um conjunto de critérios de qualidade de dados foi aplicado para avaliar a completude, precisão e relevância dos dados integrados.

Auditoria de Dados: Foi realizada uma auditoria de dados para garantir que o processo de limpeza e integração fosse concluído com êxito, revisando logs e relatórios de processamento do Apache Hop e do Jenkins.

5.2.4 Documentação e Registro:

Documentação do Processo: Todas as etapas de integração e limpeza dos dados foram devidamente documentadas, criando um registro auditável do processo.

Registro de Metadados: Metadados descritivos foram criados e excluídos para cada conjunto de dados, fornecendo informações sobre a origem, dados de coleta e quaisquer transformações aplicadas.

**5.3 Construção de modelos e visualizações de dados**

Com os dados devidamente integrados e limpos, a fase seguinte da execução do projeto centrou-se na construção de modelos de dados robustos e na criação de visualizações que transformam esses dados em insights compreensíveis e acionáveis.

Modelagem de Dados:

Estruturação de Tabelas e Relacionamentos: Com base na organização do SQL Server, desenvolvemos um modelo de dados no Power BI que reflete a estrutura relacional das tabelas e estabelece relações lógicas entre elas, como chaves primárias e estrangeiras, para garantir a integridade dos dados.

Criação de Medidas e Cálculos: Utilizamos a linguagem DAX (Data Analysis Expressions) no Power BI para criar medidas calculadas, como total acumulado de casos, taxas de crescimento e médias móveis, que são fundamentais para a análise de tendências ao longo do tempo.

Otimização para Desempenho: Otimizamos o modelo de dados para melhorar o desempenho das visualizações, usando técnicas como importação de dados versus conexões diretas, dependendo do tamanho do conjunto de dados e da necessidade de atualização em tempo real.

Desenvolvimento de Visualizações de Dados:

Seleção de visualizações: Escolhemos os tipos de visualizações que melhor representam os padrões e tendências nos dados da COVID-19, como gráficos de linha para a evolução temporal e mapas para a distribuição geográfica.

Design e Layout: Adotamos princípios de design de visualização de dados, como o uso de núcleos consistentes e legendas claras, para garantir que as visualizações sejam intuitivas e facilmente interpretáveis.

Criação de Dashboards Interativos: Incorporamos funcionalidades interativas, como filtros e segmentações, para permitir que os usuários explorem os dados de forma dinâmica e descubram insights por conta própria.

Teste e Validação das Visualizações:

Revisão com Stakeholders: Apresentamos os modelos e visualizações aos stakeholders para obter feedback, garantindo que as informações apresentadas atendam às suas necessidades e expectativas.

Ajustes e Melhorias: Com base no feedback recebido, fizemos ajustes nos modelos e visualizações para aprimorar a clareza e a utilidade, iterando o design conforme necessário.

Implementação e Publicação:

Configuração de Atualizações Automatizadas: Estabelecemos atualizações automáticas no serviço Power BI, garantindo que os dashboards reflitam as informações mais recentes.

Publicação e Compartilhamento: Os dashboards e relatórios foram publicados no Power BI Service, permitindo acesso remoto e compartilhamento com usuários autorizados.

A construção dos modelos e o desenvolvimento das visualizações são etapas cruciais para transformar dados brutos em informações importantes. No contexto deste projeto, essas etapas permitiram não apenas entender a situação atual da pandemia, mas também forneceram uma base para orientações e recomendações estratégicas para a gestão da crise de saúde pública.

**REFERÊNCIAS:**

MAIA, B. A importância do analytics na crise da Covid-19. 2020. Disponível em: https://www.mitsloanreview.com.br/post/a-importancia-do-analytics-na-crise-da-covid-19.

<https://hop.apache.org/manual/latest/getting-started/hop-what-is-hop.html>

(https://www.cienciaedados.com/pipelines-de-dados-com-apache-hop/): blog que fala sobre Hop. Não mencionei no trabalho, mas poderia.

GIL, ANTÔNIO CARLOS. Como Elaborar Projetos de Pesquisa. 4ª edição. São Paulo: Editora Atlas S.A., 2002.

# FRAGA, ADALBERTO. Microsoft Power BI: gráficos, banco de dados e configuração de relatórios. 1ª edição. Editora Alta Books, 2019.