UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
INE5643 - DATA WAREHOUSE

JACQUELINE CARDOZO
EVANDRO MACHADO
GUSTAVO BORGES
FABIANA SCHWARZ
RODRIGO JOAQUIM NUNES

Mapa de Covid

FLORIANÓPOLIS
2021

Sumário

1 Introdução

1.1 Motivação

Estimar a causalidade de crescimento e agravo da Pandemia COVID

1.2 Problema de análise de informação a ser resolvido

Aumentar o entendimento sobre a pandemia mundial de COVID

1.3 Público alvo

Cientistas de dados, profissionais da área da saúde, profissionais da área da comunicação e servidores do poder público.

2 Mapa de COVID

2.1 Fontes de dados

Para o presente trabalho, utilizaremos como fontes primárias as bases do Centro para Sistemas de Ciências e Engenharia (CSSE) da universidade americana Johns Hopkins, que contém dados mais completos em relação ao número de casos, mortes e recuperados sobre a Pandemia de COVID-19; utilizaremos também como principal fonte a Our World in Data que adicionalmente a base da Johns Hopkins, trás dados sobre vacinação, muito importante para responder as perguntas estratégicas relacionadas ao Mapa de Covid.

I. **CSSE at Johns Hopkins University**: Dados diários sobre o covid para estados/países, com as coordenadas geográficas do local, junto dos dados de mortos, recuperados, casos ativos, taxa de fatalidade, etc.

II. **Our World in Data**: Base com dados sobre covid, que contém dados parecidos com os da fonte CSSE at Johns Hopkins University, além de dados como IDH, expectativa de vida, e dados sobre vacinas.

A fim de agregar consultas mais interessantes referentes às perguntas estratégicas, também utilizaremos como fonte de dados secundárias as seguintes bases:

III. **Nager Date Public Holidays**: Base com os feriados públicos de cada país.

IV. **BioSpace - Preço e Eficácia das Vacinas**: Lista das vacinas disponíveis, com informações sobre preço, eficácia e tecnologia utilizada.

V. **Membros do COVAX Facility**: Lista dos países membros do COVAX Facility.

2.2 Contextualização do problema

2.2.1 Problemas e oportunidades

Estamos vivendo uma das maiores catástrofes da nossa geração. A pandemia do vírus COVID-19 é um problema global cujo impacto econômico e social ainda estamos tentando entender. Existem instituições renomadas como a universidade Johns Hopkins que disponibilizam bases abertas com dados atualizados sobre número de casos, mortes, vacinados entre outras coisas.

A diversidade de dados disponíveis é considerável, no entanto não há um sistema unificado onde pesquisadores e membros do poder público possam coletar e compilar consultas mais aprofundadas com o objetivo de aumentar o entendimento sobre a pandemia de COVID-19.

Para criar uma base unificada mais robusta seria necessário a mão de obra qualificada, com conhecimento técnico para angariar os dados das diversas fontes disponíveis e criar a estrutura necessária para realizar tais consultas. Para tal, o presente trabalho mostra-se como oportunidade de concretizar esse projeto, na esperança de entender melhor como cada país está lidando com os fatores envolvidos na Pandemia e traçar uma régua para conseguir o padrão mínimo de igualdade para comparativo.

2.2.2 Objetivos

Existem diversos dashboards e websites de veículos de comunicação no ambiente online que disponibilizam dados sobre número de casos de COVID, mortes, recuperados, vacinados e afins. Entretanto, poucos desses meios promovem um aprofundamento na análise da grande quantidade de dados que hoje temos disponíveis sobre a pandemia de COVID. Portanto, o objetivo do nosso projeto de Data Warehouse é realizar o aprofundamento desses dados para estimar a causalidade do crescimento e agravo da Pandemia de COVID.

2.2.3 Perguntas estratégicas

Além das consultas que outros sistemas já possibilitam, pretendemos também, com a agregação de outros dados, permitir a análise de questões como:

1. Quanto o desenvolvimento dos países (IDH) tem a ver com o número de mortes?
2. Quanto a densidade demográfica tem a ver com o número de mortes?
3. Quanto países mais ou menos desenvolvidos estão conseguindo imunizar sua população?
4. Quanto ser ou não membro da COVAX está impactando a faixa de percentual de imunizados de um país?
5. Quanto tempo leva desde as primeiras imunizações até uma queda relevante nas mortes?
6. Qual continente possui maior taxa de casos?
7. Qual continente possui maior taxa de mortalidade?

2.3 Escopo

2.3.1 Justificativa

No contexto da disciplina de Data Warehouse, o presente grupo se propôs a realizar um projeto que seja de utilidade para o público em geral.

Atualmente, estamos vivendo em um momento de pandemia, onde o tema COVID-19 é muito presente no dia a dia de todos. Sendo assim, propusemos a criação de um sistema que possibilite realizar análises sobre a pandemia. Existem vários sistemas analíticos que possibilitam o acesso de forma rápida e fácil a dados sobre a pandemia, porém notamos que as perguntas que nós evidenciamos (seção 2.2.3) não são respondidas por esses sistemas existentes. Notamos também que os dados que tornaria possível responder a essas perguntas existem, porém por diversas vezes eles não estão dispostos de uma forma integrada, de simples manuseio.

Acreditamos que as perguntas que nós destacamos são perguntas úteis para obter um panorama mais geral sobre como os países estão lidando com a pandemia, e também obter um *status* desses países em relação a situação global. Portanto o desenvolvimento do sistema proposto serve para responder essas perguntas, e somar no conhecimento geral sobre a pandemia.

2.3.2 Riscos

Os riscos mais evidentes que se observam na questão do desenvolvimento do presente trabalho, é o insucesso na integração das bases de dados, e como consequência disso, a impossibilidade de realizar as atividades técnicas planejadas.

As bases de dados contém dados que são possíveis de serem relacionados, porém esses dados estão separados em diversas bases. O risco da integração dessas bases vem do fato de ser impossível, ou inviável, realizar essas integrações desses dados disponíveis. Junto com isso, as atividades técnicas ficam comprometidas.

A também a possibilidade da realização das atividades técnicas dentro do prazo do trabalho ser impossibilitada, por conta de uma integração que possa vir a ser feita de uma forma indevida (ineficiente), de modo que a implementação do sistema fique comprometida por conta dessa integração mal formada.

2.3.1 Exclusões

Não serão abordados detalhes de cidades e estados pois há uma dificuldade muito grande em encontrar esse tipo de dado granular a um nível mundial. Optamos por nos atermos somente a países, com a possibilidade apenas de subir o nível para continentes.

Devido à complexidade em termos de modelagem, optou-se por, neste primeiro momento, manter de fora detalhes de clima e eventos (feriados, shows, eventos religiosos, eventos políticos, qualquer coisa que possa causar aglomerações) por região. Por se tratar de dados que variam, obrigatoriamente, com a geografia (cidade/estado/país) e com o tempo (data específica), este mapeamento se torna muito complexo para ser modelado no tempo da disciplina, mas fica como uma análise muito relevante para uma continuação do projeto no futuro, se for o caso.

Outro motivo que nos levou a excluir uma análise por característica climática (mesmo que não fosse dependente da data) foi a decisão de tirar a visão por cidade ou estado. Como existem países como o Brasil, que se estendem por uma latitude muito longa (+5° e -33°), não faria sentido tomar um destes pontos como ponto representativo do país inteiro.

Existem dados de vacinação que não deixam claro qual foi a vacina utilizada em um ou outro país. Desta forma, também não é esperado que todos os países e/ou continentes venham a ter todos os detalhes de tipo de vacina (e custos delas).

A análise por continentes não entrará em nível detalhado de IDH porque não existe a métrica de IDH por continente.

2.3.2 Fatores críticos de sucesso

- Unificar fonte de acesso a informações para a análise relativa aos impactos da pandemia em diversas abordagens possíveis;
- Simplificar o entendimento dos impactos listados acima;
- Tornar-se referência na obtenção de informações ao público interessado.

2.4 Plano de desenvolvimento

2.4.1 Cronograma

A1: Extrair fontes de dados disponíveis em CSV.
A2: Extrair e compilar os dados de Preço e Eficácia das Vacinas (BioSpace, 2021) em formato CSV.
A3: Extrair e compilar dados sobre Membros do COVAX Facility (WHO, 2020) em formato CSV.
A4: Validar e limpar dados.
A5: Definir a ferramenta para desenvolvimento do projeto de Data Warehouse.
A6: Integrar dados na ferramenta definida.
A7: Transformar os dados de acordo com os requisitos técnicos do projeto.
A8: Carregar os dados transformados para a conclusão da nova base de dados.
A9: Validar a nova base com realização de consultas propostas de acordo com as perguntas técnicas.
A10: Entrega e apresentação do projeto.

Etapas	**2021**						
	Mar-29	**Apr-05**	**Apr-12**	**Apr-19**	**Apr-26**	**May-03**	**May-10**
A1	**X**	**X**					
A2	**X**	**X**					
A3	**X**	**X**					
A4		**X**	**X**				
A5			**X**				
A6				**X**			
A7				**X**	**X**		
A8					**X**	**X**	
A9						**X**	
A10							**X**

2.4.2 Equipe

Todos os membros da equipe são responsáveis por todas as etapas do processo. Estamos nos reunindo para realizar o levantamento de fontes, a criação da modelagem, teorias e perguntas a serem feitas. Da mesma forma, toda a parte de ETL e criação do data warehouse final será realizada em grupo.

Caso realizado o “front” com os dados, serão as responsáveis principais Jacqueline e Fabiana.

2.4.3 Custos

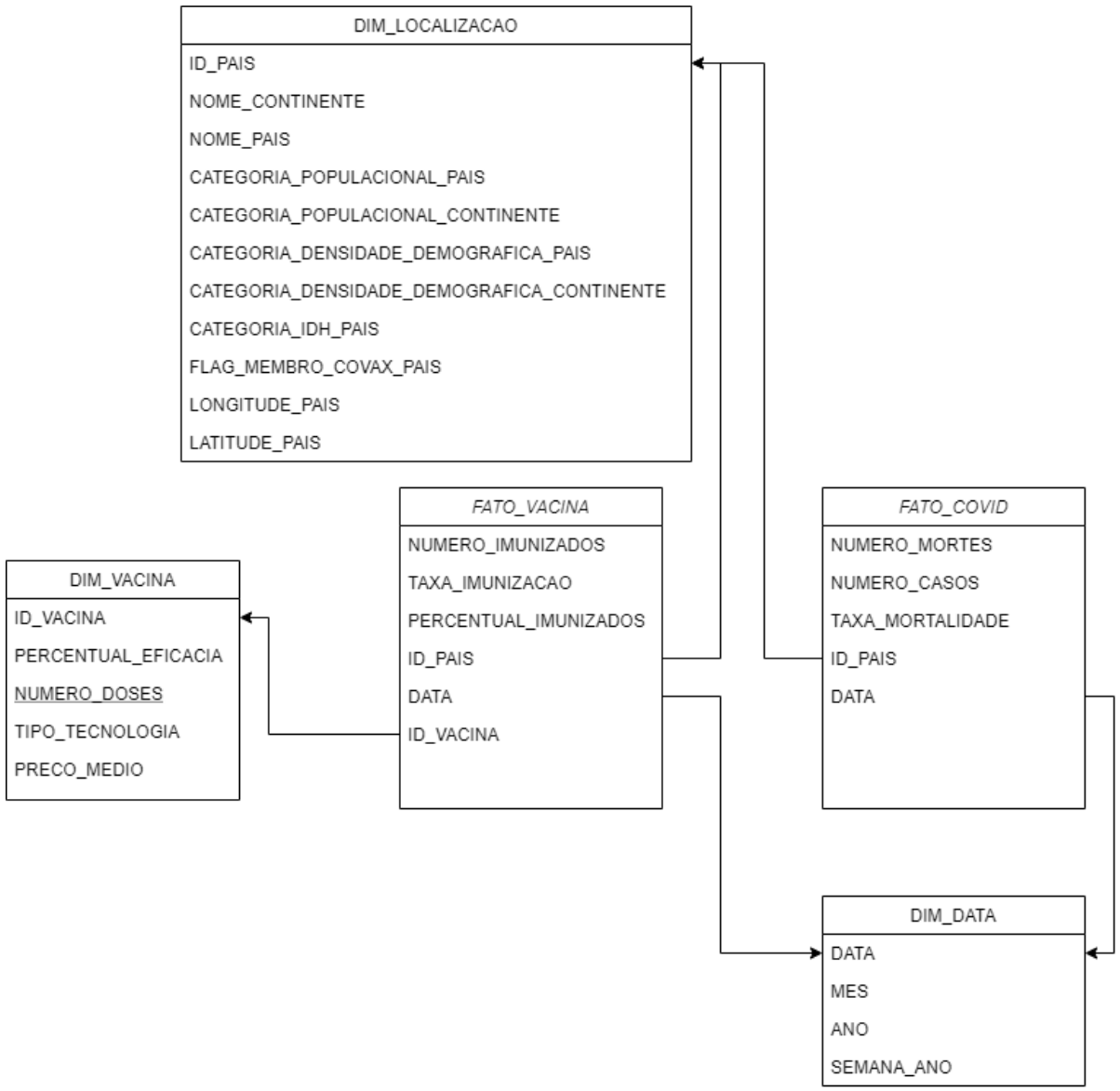
- Pagamento de pró-labore de 8.000,00 por componente da equipe pelo trabalho executado;
  Valor do item: 40.000,00
- Aquisição de notebook com elevada capacidade de processamento para execução do projeto (Notebook Acer ConceptD 3 CN315-71P-5527 Intel Core i5 12GB NVIDIA Quadro T1000 256GB SSD Windows Pro)
  Valor R$ 11.099,00

Valor total estimado para o projeto: 51.099,00

2.5 Modelagem dimensional

O esquema dimensional inicialmente proposto, acompanhado de esquema de tabelas e dicionário de dados.

2.5.1 Esquema de tabelas



2.5.2 Dicionário de dados

Tabela: FATO_COVID
Descrição geral: traz dados granulares de número de mortes.

Dado	Tipo Dado	Descrição
NUMERO_MORTES	INTEIRO (NAO NULO)	Número de mortes oficiais registradas por dia.
NUMERO_CASOS	INTEIRO (NAO NULO)	Número de casos oficiais (testados) registrados por dia.
TAXA_MORTALIDADE	DECIMAL (NAO NULO)	Número de mortes por 1 milhão de habitantes.

Tabela: FATO_VACINA
Descrição geral: traz dados granulares de número de imunizados.

Dado	Tipo Dado	Descrição
NUMERO_IMUNIZADOS	INTEIRO (NAO NULO)	Número de pessoas completamente vacinadas no país. Caso a vacina seja 2 ou mais doses, só será considerada a última dose na contagem.
PERCENTUAL_IMUNIZADOS	DECIMAL (NAO NULO)	Percentual de pessoas completamente vacinadas no país. Caso a vacina seja 2 ou mais doses, só será considerada a última dose na contagem.
TAXA_IMUNIZACAO	DECIMAL (NAO NULO)	Número de imunizados por 1 milhão de habitantes. Caso a vacina seja 2 ou mais doses, só será considerada a última dose na contagem.

Tabela: DIM_VACINA
Descrição geral: traz dados qualitativos e quantitativos sobre cada tipo de vacina.

Dado	Tipo Dado	Descrição
PERCENTUAL_EFICACIA	DECIMAL (NAO NULO)	Eficácia percentual da vacina.
NUMERO_DOSES	INTEIRO (NAO NULO)	Número de doses necessárias para imunização.
TIPO_TECNOLOGIA	CARACTERES (NULO)	Tipo de tecnologia utilizada na vacina (mRNA, adenovírus etc.).

PRECO_MEDIO	DECIMAL (NULO)	Preço médio por vacina.

Tabela: DIM_DATA
Descrição geral: traz a data e possíveis desdobramentos para drill down.

Dado	Tipo Dado	Descrição
DATA	DATA (NAO NULO)	Data informando dia/mês/ano.
MES	INTEIRO (NAO NULO)	Mês da data informada.
ANO	INTEIRO (NAO NULO)	Ano da data informada.
SEMANA_ANO	INTEIRO (NAO NULO)	Traz a semana da data atual em relação ao início do ano.

Tabela: DIM_LOCALIZACAO
Descrição geral: traz dados intrínsecos a países.

Dado	Tipo Dado	Descrição
NOME_PAIS	CARACTERES (NAO NULO)	Nome do país.
NOME_CONTINENTE	CARACTERES (NAO NULO)	Nome do continente
CATEGORIA_POPULACIONAL_PAIS	CATEGORIA (NAO NULO)	Faixas que definem um tamanho para cada país: ● Pequeno: até 100 mil ● Pequeno-médio: até 1 milhão ● Médio: até 100 milhões ● Grande: até 1 bilhão ● Gigantesco: mais que 1 bilhão
CATEGORIA_POPULACIONAL_CONTINENTE	CATEGORIA (NAO NULO)	Faixas que definem um tamanho para cada país: ● Pequeno: até 100 milhões ● Pequeno-médio: até 1 bilhão ● Médio: até 3 bilhões ● Grande: até 3 bilhões
CATEGORIA_DENSIDADE_DEMOGRAFICA_PAIS	CATEGORIA (NAO NULO)	● Pouco populoso: até 10 ● Levemente populoso: até 50 ● Populoso: até 100 ● Muito populoso: até 1.000

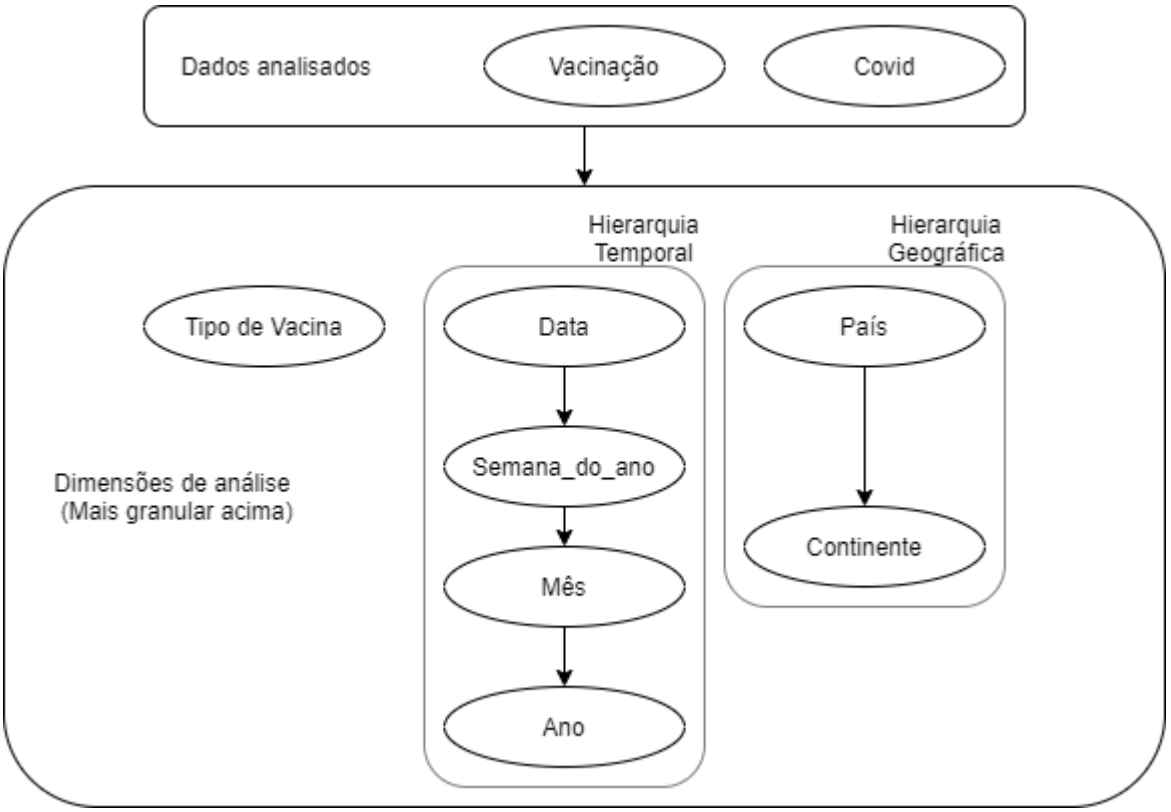
		• Extremamente populoso: mais que 1.000
CATEGORIA_DENSIDADE_DEMOGRAFICA_CONTINENTE	CATEGORIA (NAO NULO)	• Pouco populoso: até 50 • Levemente populoso: até 100 • Populoso: até 200 • Muito populoso: mais que 200
CATEGORIA_IDH_PAIS	CATEGORIA (NAO NULO)	Índice de desenvolvimento humano • Baixo: 0.350 – 0.549 • Médio: 0.550 – 0.699 • Alto: 0.700 – 0.799 • Muito alto: 0.800 – 1.000
FLAG_MEMBRO_COVAX_PAIS	BOOLEANO (NAO NULO)	Booleano para saber se o país faz parte do Consórcio mundial para compra de vacinas
LATITUDE_PAIS	DECIMAL (NAO NULO)	Latitude do país
LONGITUDE_PAIS	DECIMAL (NAO NULO)	Longitude do país

2.6 Hierarquias

Trabalhamos com duas hierarquias principais:

1) Hierarquia temporal: Propomos drill dentro do ano para análise mensal e semanal. Também, outra possibilidade seria a de comparativo direto entre mesmas semanas de anos diferentes, ou mesmos meses em anos diferentes.
2) Hierarquia geográfica: Propomos drill up dos países, fazendo uma análise mais macro, a partir de continentes, depois descendo para países.

A figura abaixo mostra, dentro das dimensões de análise, as hierarquias citadas acima.



3 Conclusões Parciais

Com o presente trabalho, podemos concluir que a pandemia do vírus COVID-19 é um problema global cujo impacto econômico e social ainda necessita ser analisado. Há disponível na internet, diversos dashboards e websites de veículos de comunicação no com dados sobre número de casos, mortes, vacinados etc., porém poucos deles permitem aprofundamento na análise da grande quantidade de dados que hoje temos disponíveis sobre a pandemia de COVID. Diante disso, constitui-se a oportunidade de usar nossos conhecimentos como alunos da disciplina de Data Warehouse para extrair, transformar e carregar as diversas bases de dados disponíveis sobre os número da pandemia e usá-los de maneira a responder perguntas mais otimizadas sobre causalidade de crescimento e agravo da Pandemia COVID-19.

Entre as importantes questões que podem ser respondidas através da integração das fontes de dados especificadas no presente trabalho, consideramos importante relacionar o Índice de Desenvolvimento Humano com o número de mortes e vacinados, assim como identificar o impacto do consórcio COVAX Facility na taxa de mortalidade e dados de imunização. Consideramos ser interesse do público em geral também entender a partir de que momento a taxa de imunização impacta na redução das mortes por COVID.

Consideramos os riscos associados a usar uma grande gama de dados e possibilidade de insucesso na integração das bases de dados. Apesar dos desafios, é de entendimento dos membros do grupo que as questões técnicas especificadas são muito importantes, servindo de fator motivacional para a realização das tarefas especificadas no cronograma.