SCC0245
Processamento Analítico de Dados

Prof.ª Dr.ª Cristina Dutra de Aguiar
PAE João Pedro de Carvalho Castro

Trabalho prático 2
Dados sobre saúde e clima no Brasil

Alunos:

- Dalton Hiroshi Sato nUSP 11275172
- Lucas Yuji Matubara nUSP 10734432
- Savio Duarte Fontes nUSP 10737251

Sumário

Parte 1 - Correção

Descrição do problema

Tema

Com uso de bases de dados provenientes do kaggle e do SUS, nosso tema é sobre a saúde no Brasil, em específico a partir do ano de 2020, com uso de informações sobre relatórios de doenças respiratórias e informações sobre variações do clima e do tempo no Brasil.

Como o mundo está passando por um processo de transição epidemiológica, no qual algumas atividades estão retornando à normalidade, seria de grande interesse conhecer sua evolução e possivelmente detectar algumas características e peculiaridades.

Processos de negócios

O DataWarehouse (DW) possuirá foco em Clima e sua influência nas Doenças Respiratórias, em especial, no Sars-Cov-2.

Por meio dos datasets do Opendatasus (Datasus) e do Kaggle (INMET) os integrantes farão o ETL (Extract, Transform, Load) pelo Pentaho Data Integration, para tratar os dados obtidos.

Essas informações trabalhadas seriam importantes a profissionais de saúde como: médicos, enfermeiros, técnicos em enfermagem, farmacêuticos, o que facilitaria o atendimento dos profissionais disponibilizando uma análise anterior à consulta, também ajudaria no controle das ações governamentais no isolamento social.

Serão feitas buscas quando os profissionais de saúde acharem necessário informações quanto às dimensões trabalhadas para aumentar a produtividade em seus trabalhos e se atualizarem sobre o cenário atual.

Fatos:

- Saúde
    - Atributos:
        - Registro Ocorrência (Aditivo, por meio de SUM, de modo a contabilizar a quantidade de entradas na tabela de fatos)
        - Gestante (Aditivo, por meio de SUM, de modo a contabilizar a quantidade de entradas na tabela de fatos)
    - Dimensões:
        - Local
        - Estado Paciente
        - Paciente
        - Data
        - Grupo Sintomas
        - Grupo Doenças

- Clima:
    - Atributos:
        - Precipitação Total (Aditivo, por meio de SUM para calcular a precipitação total, mediante um filtro por data e/ou região)
        - Temp Máx (Não aditivo)
        - Temp Mín (Não aditivo)
        - Umidade Máx (Não aditivo)
        - Umidade Mín (Não aditivo)
        - Vento Máx (Não aditivo)
    - Dimensões:
        - Local
        - Data

Dimensões:

- Local - Clima
    - Estado - Da admissão do paciente e da medição do clima
    - Região - Da admissão do paciente e da medição do clima
    - País - Brasil
    - Estação - Estação de medição
    - CódigoEstação - Código da estação
    - Relacionamento:
        - All > país > região > estado > estação
- Local - Saúde
    - Unidade - Hospital de entrada
    - Município - Cidade de entrada
    - Estado - Estado de entrada
    - Região - Região de entrada
    - MunicípioNúmero - Código do município
    - RegiãoNúmero - Codigo da região
    - UnidadeNúmero - Código da região
    - Relacionamento
        - All > região > Estado > Município > Unidade
- Situação Paciente
    - Estado doente - Se o paciente já foi diagnosticado com alguma doença
    - Estado óbito - Se o paciente foi a óbito, entre data de admissão e data da formação do relatório
- Paciente
    - Nome - Do paciente (Removido)
    - CPF - Do paciente (Removido)
    - RG - Do paciente (Removido)
    - Data de nascimento - Do paciente
    - Sexo - Do paciente
    - TomouVacina - Se tomou vacina
    - Data1Dose - Data da primeira dose
    - Fab1 - Fabricante da primeira dose
    - Data2Dose - Data da segunda dose

    - Fab2 - Fabricante da segunda dose
    - DataRef - Data da dose de reforço
    - FabRef - Fabricante da dose de reforço
- Data
    - Data Completa - data em formato DD/MM/AAAA
    - Dia - o dia da data
    - Semana - a semana em relação ao ano no momento
    - Mês - o mês da data
    - Mês-ano - o mês e o ano
    - Mês nome - nome do mês
    - Trimestre - o trimestre em relação ao ano no momento
    - Trimestre-ano - o trimestre em relação ao ano no momento, e o ano
    - Semestre - o semestre em relação ao ano no momento
    - Semestre-ano - o semestre em relação ao ano no momento, e o ano
    - Ano - o ano da data
    - Relacionamento:
        - ALL > ano > semestre > trimestre > mês > semana > dia
- Sintomas
    - Chave sintoma - chave do sintoma
    - Sintoma - Nome do sintoma
- Ponte Grupo Sintoma
    - Grupo sintoma - Grupo de sintomas apresentados pelo paciente
    - Sintoma - Chave do sintoma
    - Fator de ponderação - Auxilia na construção de somatórios
- Doença
    - Chave doença - chave da doença
    - Doença - Nome da doença
- Ponte Grupo Doença
    - Grupo doença - Grupo de doenças apresentados pelo paciente
    - Doença - Chave da doença
    - Fator de ponderação - Auxilia na construção de somatórios
- Vacinas (Removido)
    - ID vacina - Chave da vacina
    - Nome_empresa - Nome da empresa que produz a vacina
- Dose (Removido)
    - ID dose - Chave da dose
    - Número Dose - Número da dose tomada

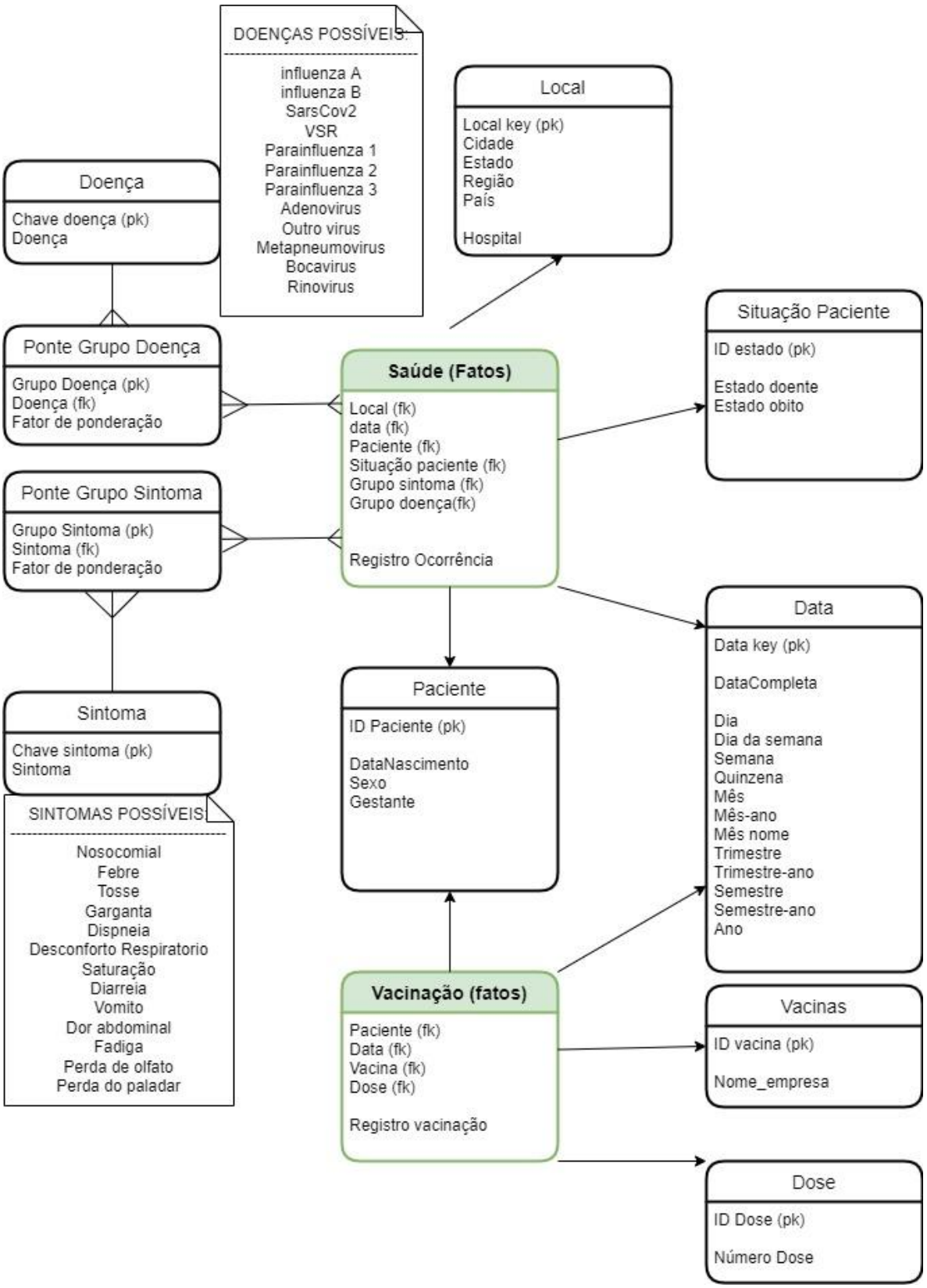
Povoamento:

- Dimensão Data será povoada por meio de programação
- Dimensão Local será povoada por meio da integração das nossas duas bases de dados, o DataSus e o Kaggle, pegando os atributos referentes à Dimensão Local
- Dimensão Paciente será povoada utilizando os dados referentes aos atributos da Dimensão Paciente que estão na base de dados DataSus
- Dimensão Estado Paciente será povoada utilizando os dados da base do dataSus referentes se ele está doente e se ele veio a óbito

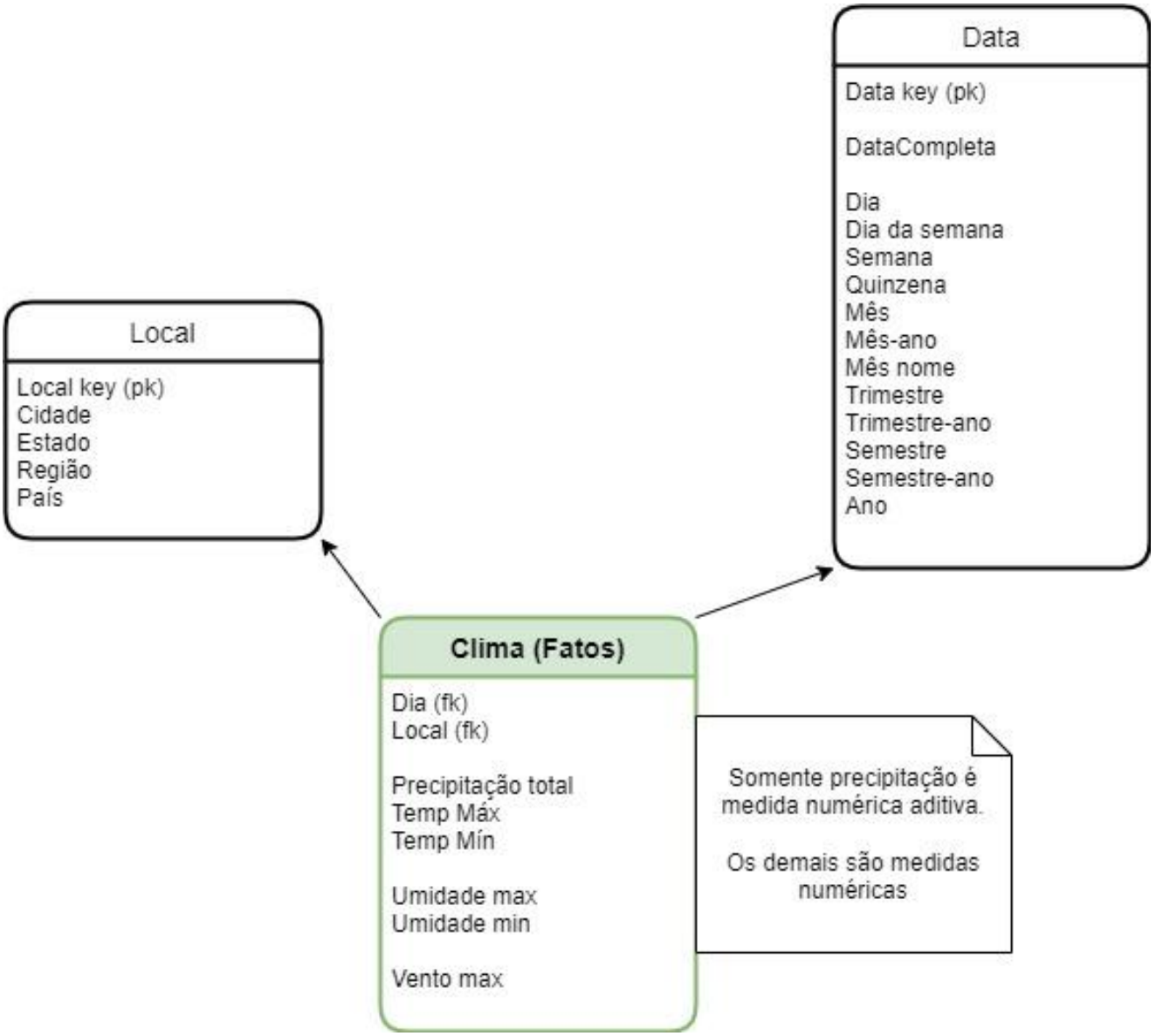
- Dimensão Vacina será povoada utilizando os nomes das vacinas que se encontram na base de dados do DataSus (Removido)
- Dimensão Grupo Doenças será povoada utilizando os dados da base do dataSus de forma que todos as doenças sejam colocadas na tabela de doenças e os grupos de doenças que os pacientes têm serão colocados na tabela Grupo Doenças
- Dimensão Grupo Sintomas será povoada utilizando os dados da base do dataSus de forma que todos os sintomas sejam colocadas na tabela de sintomas e os grupos de sintomas que os pacientes têm serão colocados na tabela Grupo Sintomas
- Fato Registro Ocorrência recebe um inteiro de valor 1 para fazermos operações aditivas em cima da tabela de fatos
- Fato Gestante recebe um inteiro de valor 1, caso gestante e 0 caso contrário, para fazermos operações aditivas em cima da tabela de fatos
- Fato Precipitação total será povoado utilizando a soma diária das precipitações que estão na base de dados do Kaggle
- Temp Máx será povoado utilizando a temperatura máxima diária que está na base de dados do Kaggle
- Temp Mín será povoado utilizando a temperatura mínima diária que está na base de dados do Kaggle
- Umidade Máx será povoado utilizando a umidade máxima diária que está na base de dados do Kaggle
- Umidade Mín será povoado utilizando a umidade mínima diária que está na base de dados do Kaggle
- Vento Máx será povoado utilizando o medição máxima diária do vento que está na base de dados do Kaggle
- Os atributos da tabela de fatos Clima, exceto precipitação total, são apenas medidas numéricas, e decidimos que não seria necessário e eficiente criar dimensões diferentes para cada uma delas.
- Caso seja observado que os dados extraídos não sejam suficientes para uma análise, e que as consultas não apresentem uma quantidade de dados suficientes, dados sintéticos para os esquemas de Vacinação e de Saúde podem ser inseridos.

Esquemas

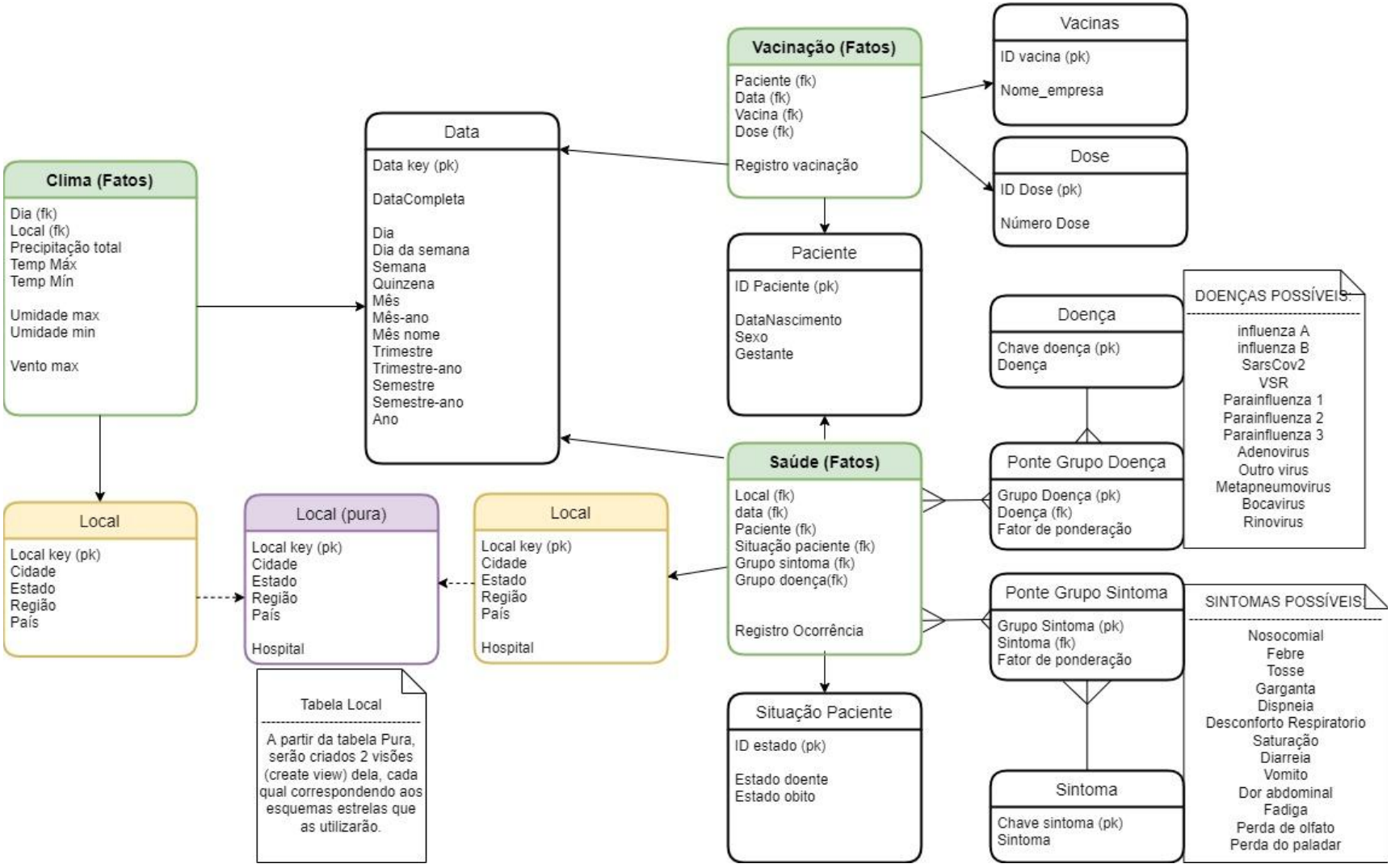
Esquema Estrela 1 - Atualizado



Esquema Estrela 2 - Atualizado



Constelação de fatos - Atualizado



A constelação de fatos foi projetada de modo que nossos 3 esquemas estrelas (Saúde, Vacinação e Clima) pudessem se interligar. No projeto, Clima se relaciona com Saúde de acordo com a data e com o local, de modo que seja possível fazer uma associação de admissões em hospitais de acordo com a condição climática em certas regiões (ar seco coopera com aparição de alguns sintomas e agravantes de algumas doenças respiratórias).

Ao mesmo tempo, Saúde se relaciona com Vacinação, de modo que seja possível verificar a ocorrência de certas doenças de acordo com o grau de vacinação, em específico, o SarsCov-2.

Consultas

- **Slice and Dice**: Verificar entradas mensais em hospital que aconteceram na cidade de São Paulo no ano de 2021
- **Drill-Across**: Verificar o número de infectados (qualquer vírus) e de precipitação diários durante os 4 primeiros meses de 2021;
- **Roll-Up**: Verificar precipitação total por mês por região.
- **Drill-Down**: Verificar por semana quantos foram os gestantes ao longo do tempo
- **Pivot**: Verificar número de contágios por semana por estado, trocando as perspectivas de tempo e local.

Parte 2

Transformações de dados

Da base de dados referentes ao clima, foram mantidas somente as entradas posteriores ao dia 1º de Janeiro de 2021, início do período da base de dados do SUS. Mantendo as colunas que se referem à Precipitação total, Temperatura Máxima e Mínima, Umidade Máxima e Mínima e Vento. Após isso, como os dados foram gerados a cada hora, foram compilados e resumidos em um espaço de tempo por dia, calculando a soma da precipitação do dia, e os valores máximos e mínimos de temperatura, umidade e vento (somente máximo).

No final, dos 10 Gb iniciais, restaram cerca de 2.5 Mb de dados de todas as 622 estações do Brasil, nas 5 regiões. Porém, foi verificado que os dados meteorológicos obtidos continham informações até o mês de abril de 2021, e por essa razão as consultas que envolvem o clima foram limitados dentro deste período.

A base de dados de saúde foi retirada do SUS e consistia em 2 arquivos separados, totalizando 1.2 Gb de dados. Foi observado que dados como Nome, CPF, RG dos pacientes foram omitidos do arquivo publicado, em função da LGPD. Isso fez com que a dimensão de pacientes não pudesse ser trabalhada da maneira esperada, então foi preenchida como se cada paciente fosse um novo paciente.

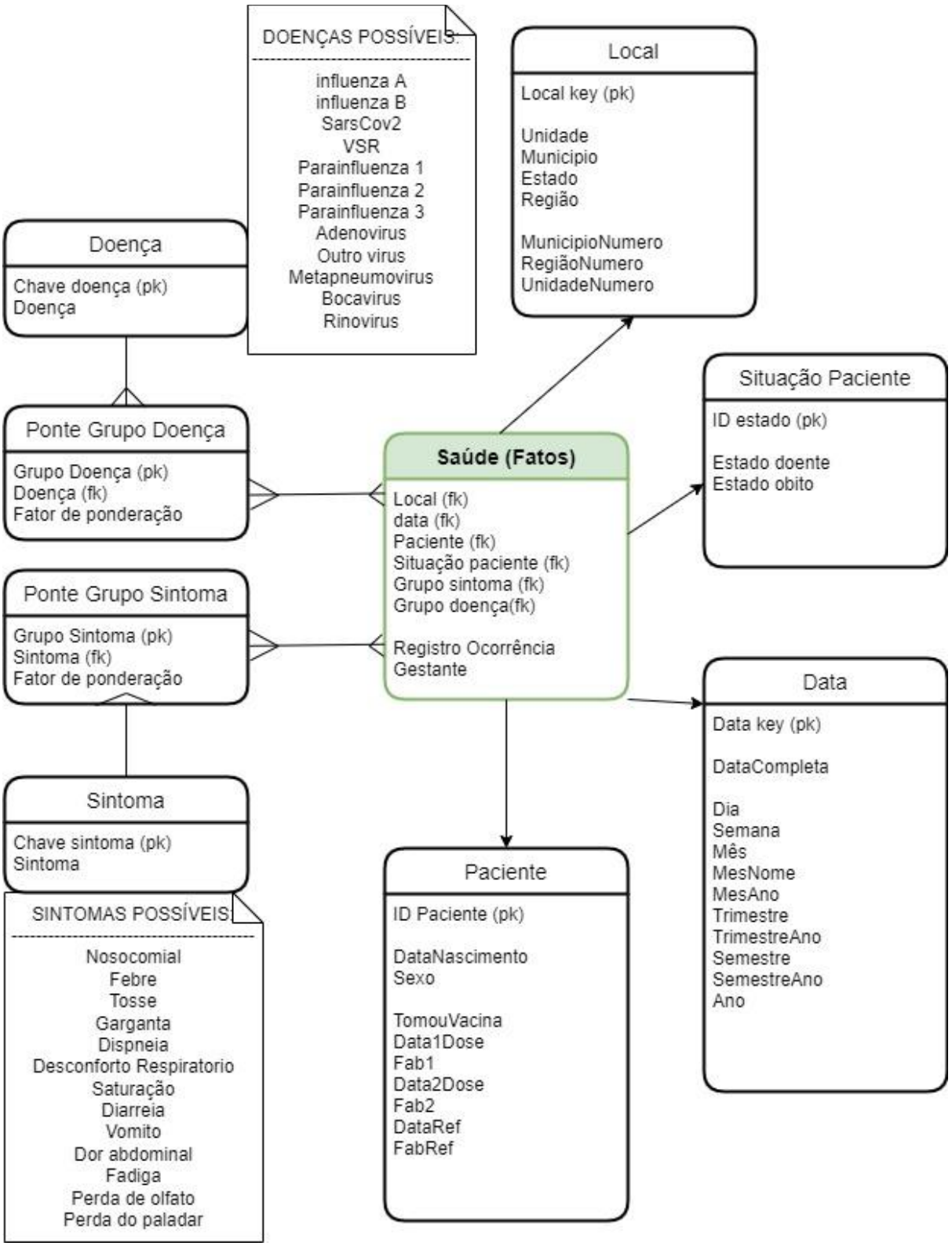
Filtrando as colunas CS_SEXO, DT_NASC, CS_GESTANT, NOSOCOMIAL, FEBRE, TOSSE, GARGANTA, DISPNEIA, DESC_RESP, SATURACAO, DIARREIA, VOMITO, DOR_ABD, FADIGA, PERD_OLFT, PERD_PALA, influenza_A, influenza_B, VSR, SarsCov2, Parainfluenza_1, Parainfluenza_2, Parainfluenza_3, Parainfluenza_4, Adenovirus, Metapneumovirus, Bocavirus, Rinovirus, Outro_virus, SG_UF_NOT, ID_REGIONA, CO_REGIONA, ID_MUNICIP, CO_MUN_NOT, ID_UNIDADE, CO_UNI_NOT, VACINA_COV; DOSE_1_COV; DOSE_2_COV; DOSE_REF; FAB_COV_1; FAB_COV_2; FAB_COVREF, foram criados diversos arquivos .csv que condizem com as dimensões propostas e as tabelas de fatos.

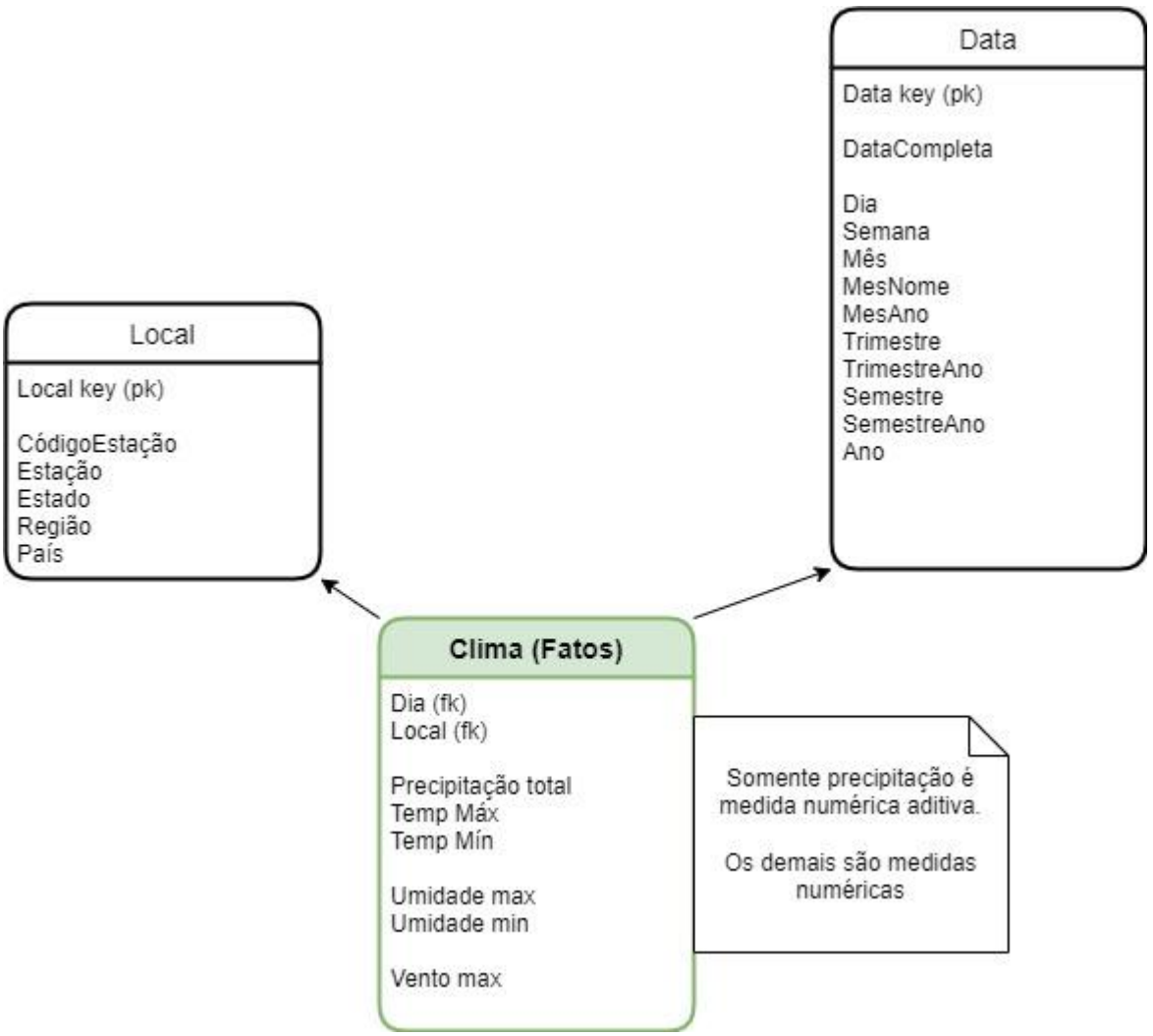
Porém, ao analisar erros de retorno de filtragem com uso do Python 3 e Pandas, percebemos que alguns dados não poderiam ser utilizados de maneira útil, e isso acarretaria em mudanças em nossos esquemas estrela e, consequentemente, na constelação de fatos.

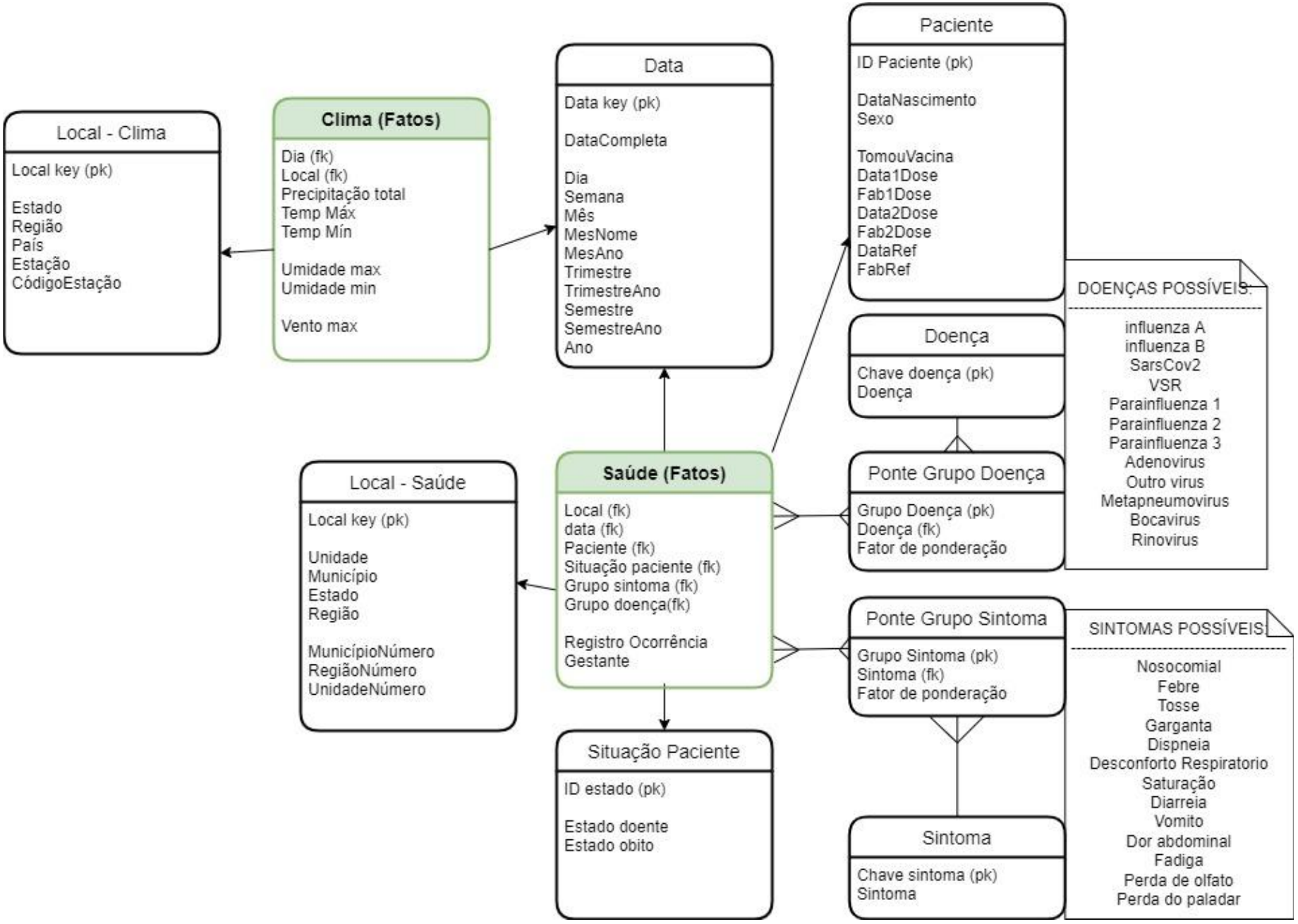
Os campos sobre a vacinação não foram preenchidos corretamente pelos usuários do sistema, e apesar da existência de um dicionário de dados e de uma documentação sobre as funcionalidades de cada campo, não havia um padrão no preenchimento dessas colunas, gerando uma grande confusão aos seus valores e que demandariam grande esforço e tempo para tentar filtrá-las adequadamente. Por essa razão, as dimensões relativas à vacinação foram incorporadas na dimensão paciente.

As dimensões de Sintomas, Doenças e suas pontes se mostraram um grande desafio para que nós pudéssemos compreender seu funcionamento. No fim, conseguimos filtrar os dados e preenchê-las, mas o resultado final cumpre tanto seu propósito quanto o propósito de uma tabela bugiganga, reflexo da falta da experiência com esse tipo de tarefa.

Sobre a dimensão em comum Local, não havia consistentemente uma maneira de relacionar ambos os esquemas estrelas de Clima e de Saúde. Isso ocorre em função de diferentes referências que ambas as fontes de dados continha: o Clima possuía estações de medição que cobriam diversas cidades, e cidades que possuíam diversas estações. Além disso, Saúde não faziam referência à sua cidade, e perante a existência de mais de 5000 municípios no Brasil, outra tarefa extremamente trabalhosa surgiu. Desta forma e diante da situação, decidimos por mudar os esquemas estrelas e seus relacionamentos na constelação de fatos, como se segue:


DOENÇAS POSSÍVEIS:
influenza A
influenza B
SarsCov2
VSR
Parainfluenza 1
Parainfluenza 2
Parainfluenza 3
Adenovirus
Outro virus
Metapneumovirus
Bocavirus
Rinovirus
Local
Local key (pk)
Unidade
Municipio
Estado
Região
MunicipioNumero
RegiãoNumero
UnidadeNumero
Doença
Chave doença (pk)
Doença
Situação Paciente
ID estado (pk)
Estado doente
Estado obito
Ponte Grupo Doença
Grupo Doença (pk)
Doença (fk)
Fator de ponderação
Saúde (Fatos)
Local (fk)
data (fk)
Paciente (fk)
Situação paciente (fk)
Grupo sintoma (fk)
Grupo doença(fk)
Registro Ocorrência
Gestante
Ponte Grupo Sintoma
Grupo Sintoma (pk)
Sintoma (fk)
Fator de ponderação
Data
Data key (pk)
DataCompleta
Dia
Semana
Mês
MesNome
MesAno
Trimestre
TrimestreAno
Semestre
SemestreAno
Ano
Sintoma
Chave sintoma (pk)
Sintoma
Paciente
ID Paciente (pk)
DataNascimento
Sexo
TomouVacina
Data1Dose
Fab1
Data2Dose
Fab2
DataRef
FabRef
SINTOMAS POSSÍVEIS
Nosocomial
Febre
Tosse
Garganta
Dispneia
Desconforto Respiratorio
Saturação
Diarreia
Vomito
Dor abdominal
Fadiga
Perda de olfato
Perda do paladar


Data
Data key (pk)
DataCompleta
Dia
Semana
Mês
MesNome
MesAno
Trimestre
TrimestreAno
Semestre
SemestreAno
Ano
Local
Local key (pk)
CódigoEstação
Estação
Estado
Região
País
Clima (Fatos)
Dia (fk)
Local (fk)
Precipitação total
Temp Máx
Temp Mín
Umidade max
Umidade min
Vento max
Somente precipitação é medida numérica aditiva.
Os demais são medidas numéricas


Local - Clima
Local key (pk)
Estado
Região
País
Estação
CódigoEstação
Clima (Fatos)
Dia (fk)
Local (fk)
Precipitação total
Temp Máx
Temp Mín
Umidade max
Umidade min
Vento max
Data
Data key (pk)
DataCompleta
Dia
Semana
Mês
MesNome
MesAno
Trimestre
TrimestreAno
Semestre
SemestreAno
Ano
Paciente
ID Paciente (pk)
DataNascimento
Sexo
TomouVacina
Data1Dose
Fab1Dose
Data2Dose
Fab2Dose
DataRef
FabRef
DOENÇAS POSSÍVEIS:
influenza A
influenza B
SarsCov2
VSR
Parainfluenza 1
Parainfluenza 2
Parainfluenza 3
Adenovirus
Outro virus
Metapneumovirus
Bocavirus
Rinovirus
Doença
Chave doença (pk)
Doença
Ponte Grupo Doença
Grupo Doença (pk)
Doença (fk)
Fator de ponderação
Local - Saúde
Local key (pk)
Unidade
Município
Estado
Região
MunicípioNúmero
RegiãoNúmero
UnidadeNúmero
Saúde (Fatos)
Local (fk)
data (fk)
Paciente (fk)
Situação paciente (fk)
Grupo sintoma (fk)
Grupo doença(fk)
Registro Ocorrência
Gestante
Ponte Grupo Sintoma
Grupo Sintoma (pk)
Sintoma (fk)
Fator de ponderação
SINTOMAS POSSÍVEIS
Nosocomial
Febre
Tosse
Garganta
Dispneia
Desconforto Respiratorio
Saturação
Diarreia
Vomito
Dor abdominal
Fadiga
Perda de olfato
Perda do paladar
Situação Paciente
ID estado (pk)
Estado doente
Estado obito
Sintoma
Chave sintoma (pk)
Sintoma

Com uso do PostgreSQL e pgAdmin 4, as tabelas foram criadas e povoadas com os .csv criados na etapa de filtragem, com uso de Pandas e NumPy em Python 3.

Os resultados das consultas podem ser analisados nos arquivos em anexo, tanto sua forma tabular quanto em forma gráfica. As consultas em .sql também estão incluídas nos arquivos. Vale ressaltar que todas as consultas são executadas rapidamente e os resultados obtidos fazem sentido, de acordo com a visão dos membros do grupo.

Referências Bibliográficas

Datasus. “SRAG 2021 e 2022 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19 - OPENDATASUS.” *OPENDATASUS*, 2022, https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022. Accessed 18 May 2022.

INMET. “Climate Weather Surface of Brazil - Hourly.” *Kaggle*, outubro 2021, https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region. Accessed 18 May 2022.