Trabalho de Sistemas de Apoio à Decisão
Business Intelligence

Grupo Les Amis du Codage:
Celso Antonio Uliana Junior, Felipe Salles Lopes, Lucas Avanzi

RGA: 2014.1907.003-7, RGA: 2016.1907.032-4, RGA: 2016.1907.024-3

Objetivo do Processo de BI

O objetivo principal é dar suporte à decisão para um gestor de nível nacional (gestor nível federal) com base em dados climáticos, epidemiológicos em relação ao COVID-19 e infraestrutural.

Com o sistema BI será possível responder a três perguntas:

1 - A partir da visualização de um gráfico de um país já muito afetado, descobrir a partir de que momento que as mortes começaram a aumentar significativamente, destacando a capacidade total de leitos do país em questão. Para esta questão, os países selecionados foram: Brasil, Espanha, Estados Unidos e Itália; os três últimos foram escolhidos por serem os mais afetados em todo o mundo (Obs.: China não está entre as escolhidas, pois decidimos que esta não fornece dados confiáveis).

2- A partir das informações obtidas da pergunta anterior, predizer quando no Brasil as mortes poderiam aumentar significativamente com o avanço do COVID-19.

3- Comparar ocorrências de COVID-19 entre grandes cidades que têm uma temperatura mais alta e grandes cidades que têm uma temperatura mais baixa durante a pandemia para saber se o clima pode afetar a transmissão. As cidades selecionadas foram: São Paulo, Rio de Janeiro, Camberra, Buenos Aires, Miami, Nova Iorque, Los Angeles, Madrid e Roma.

Bases de Dados Utilizadas

Link para o drive com os dados e modelagem:
https://drive.google.com/drive/folders/1UASBclx21-tdtGvxKj97HvZs2mLvfEP-?usp=sharing

Leitos para Internação Brasil:

Atributos:
Uf - Unidade federativa na qual possui os leitos (ex: MS).
leitos_sus - Quantidade total de leitos para internação do Sistema Unificado de Saúde (SUS) em um estado (ex: 239).
leitos_não_sus - Quantidade total de leitos para internação que não fazem parte do SUS (ex: 137).
total_de_leitos - Quantidade total de leitos, sendo estes SUS ou não SUS (ex. 376).

Quantidade de dados: 27 linhas, um para cada unidade federativa. O objetivo aqui não é ter em mãos a capacidade de leitos por dia e sim ter a capacidade total do país para que possamos colocar uma linha.
Baixa granularidade de dados (alto detalhamento).

Leitos por 1000 habitantes:
Atributos:
Pais - Nome do país. Tipo String (ex: Brasil).
Camas de hospital por habitante (camas / 1,000 habitantes) - Leitos por 1000 habitantes. Tipo decimal (ex: 11,5).

Quantidade de dados: 178 linhas atualmente, aumentando diariamente.
Alta granularidade de dados (baixo detalhamento).

COVID-19 mundo:

Atributos:
dateRep - Qual o dia de registro desse dado no formato MM/dd/yyyy (ex: 04/27/2020).
day - Valor inteiro do dia (ex:27).
month - Valor inteiro do mês (ex: 4).
year - Valor inteiro do ano (ex: 2020).
cases - Quantidade de novos casos nesse dia (ex: 68).
deaths - Quantidade de novos óbitos nesse dia (ex: 21).
countriesAndTerritories - Nome extenso do país ou território (ex: Brazil).
geold - A sigla geográfica do país (ex: BR)
countryterritoryCode - Código geográfico ISO3 do país (ex: BRA);
popData2018 - Número inteiro que representa a população do país em 2018 (ex: 37172386).
continentExp - Continente do país em questão (ex: Asia).

Quantidade de dados: 13.000 linhas atualmente, aumentando diariamente.

Baixa granularidade de dados (alto detalhamento).

Dados Climáticos:

Os dados climáticos para cada estão separados, porém com os mesmos atributos e tipos (com exceção para os dados de Camberra nos quais a temperatura máxima e mínima são do tipo decimal). Os atributos são:

Atributos:
Data - Data referente a temperatura do dia. Tipo date (ex: 07/05/2020).
max - Temperatura máxima do dia. Tipo int (ex: 28).
min - Temperatura mínima do dia. Tipo int (ex: 18).

Além dos atributos, as bases de dados também compartilham de sua granularidade, no qual é baixa.

As bases para esses atributos são:

- buenos-aires-temperature.xls: possui 56 linhas atualmente, aumentando diariamente.
- canberra-temperature.xls: possui 118 linhas atualmente, aumentando diariamente.
- los-angeles-temperature.xls: possui 93 linhas atualmente, aumentando diariamente.
- madrid-temperature.xls: possui 62 linhas atualmente, aumentando diariamente.
- miami-temperature.xls: possui 48 linhas atualmente, aumentando diariamente.
- new-york-city-temperature.xls: possui 58 linhas atualmente, aumentando diariamente.
- rio-de-janeiro-temperature.xls: possui 53 linhas atualmente, aumentando diariamente.
- roma-temperature.xls: possui 65 linhas atualmente, aumentando diariamente.
- sao-paulo-temperature.xls: possui 63 linhas atualmente, aumentando diariamente.

Covid-19 por cidades:

us-counties.xlsx
Atributos:
date - Data de registro. Tipo date (ex: 05/07/2020).

county - Nome do condado. Tipo String (ex: Miami-dade).
state - Nome do estado. Tipo String (ex: Florida).
fips - Código de cada condado. Tipo int (ex: 23145).
cases - Quantidade de casos acumulados. Tipo int (ex: 25).
deaths - Quantidade de óbitos acumulados. Tipo int (ex: 8).

Quantidade de dados: 95.420 linhas atualmente, aumentando diariamente.
Baixa granularidade de dados (alto detalhamento).

espana_covid19_casos.xlsx
Atributos:
cod_ine - Código correspondente à comunidade autônomas. Tipo int (ex: 3).
CCAA - Nome da comunidade autônoma. Tipo String (ex: Madrid).

Quantidade de dados: 20 linhas. Nesse caso, são as colunas que aumentam, pois as datas são colunas e seus valores são a quantidade de casos acumulados.

caso-br.xlsx
Atributos:
date - Data de registro. Tipo date (ex: 07/03/2020).
state - Nome do estado. Tipo String (ex: Mato Grosso do Sul).
place_type - Aponta se os dados são relativos a cidade ou ao estado. (Curitiba).
confirmed - Quantidade de casos acumulados. Tipo int (ex: 1542).
deaths - Quantidade de óbitos acumulados. Tipo int (ex: 154).
estimated_population_2019 - População estimada àquela região. Tipo int (ex: 548275).
city_ibge_code - Código da cidade de acordo com o IBGE. Tipo int (ex: 51).
confirmed_per_100k_inhabitants - Casos confirmados por 100 mil habitantes. Tipo decimal (ex: 754,778).
death_rate - Taxa de morte. Tipo float (ex: 0,0418).

Quantidade de dados: 35.853 linhas atualmente, aumentando diariamente.
Baixa granularidade de dados (alto detalhamento).

dpc-covid19-ita-province.xlsx
Atributos:
data - Data de registro. Tipo date (ex: 04/04/2020).
codice_regione - Código da região. Tipo int (ex: 23).
denominazione_regione - Nome da região. Tipo String. (ex: Lazio).
codice_provincia - Código da província. Tipo int (ex: 058).
denominazione_provincia - Nome da província. Tipo String (ex: Roma).
sigla_provincia - Sigla da província. Tipo String (ex: RM).

lat - Latitude da província. Tipo long (ex: 4235103167).
long - Longitude da província. Tipo long (ex: 1416754574).
totale_casi - Casos acumulados. Tipo int (ex: 399).

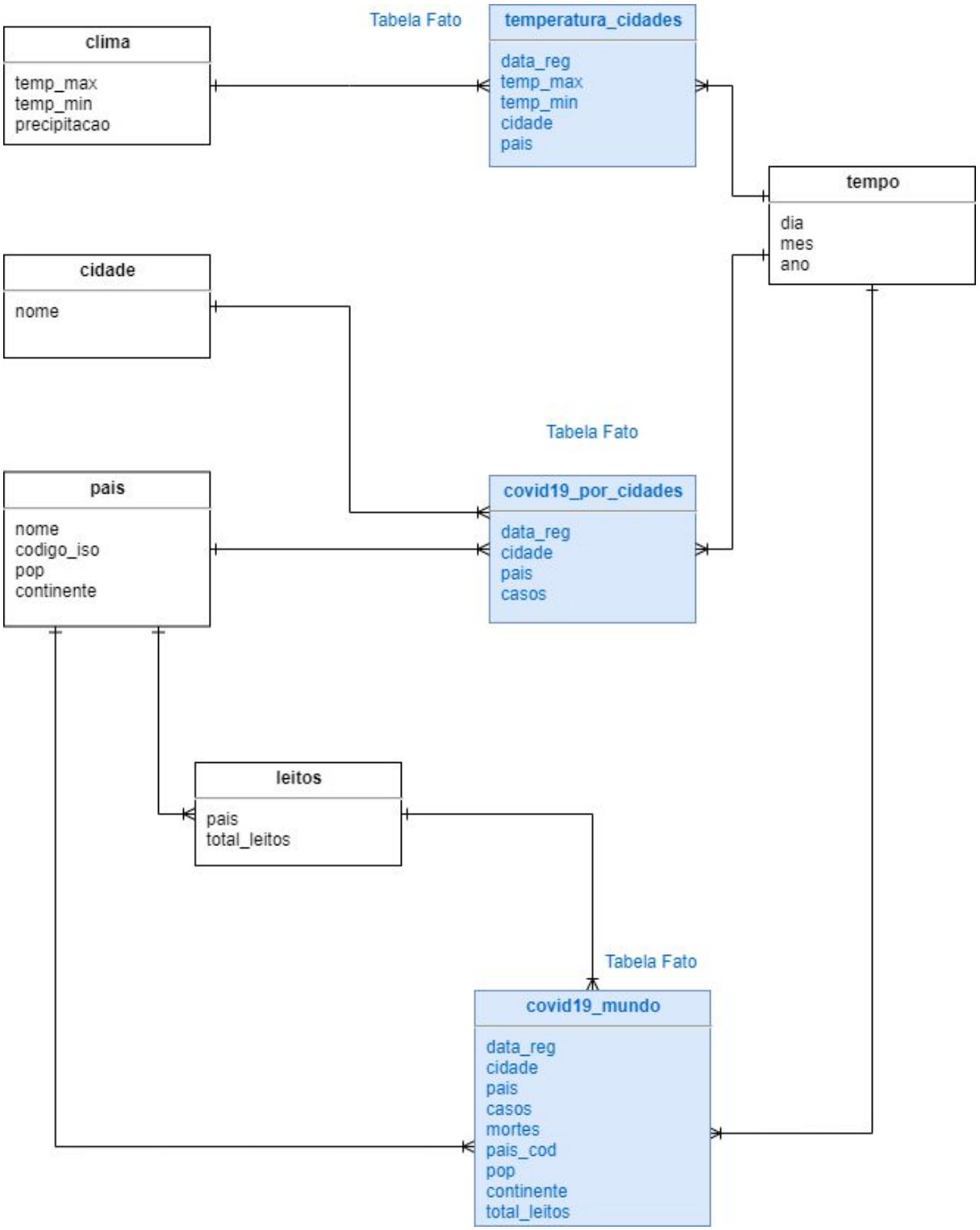
Quantidade de dados: 8.320 linhas atualmente, aumentando diariamente.
Baixa granularidade de dados (alto detalhamento).

australian-capital-territory.xlsx
Atributos:
date - Data de registro. Tipo date (ex: 07/05/2020)
territory - Território administrativo da Austrália. Tipo String (ex: ACT).
cumm_cases - Casos acumulados. Tipo int (ex: 27).

Quantidade de dados: 58 linhas atualmente, aumentando diariamente.
Baixa granularidade de dados (alto detalhamento).

Modelagem do DW

As tabelas fatos geradas foram: *temperatura_cidades, covid19_por_cidades* e *covid19_mundo.*



A modelagem do Datawarehouse pode ser encontrada nesse link:
https://app.diagrams.net/#G112HfcXuXYJF3D2bgu28rEuNJ4PNrUYGL

Obs.: Por não conseguir executar na minha máquina, não foi utilizado o SQL Power Architect e, em seu lugar, modelamos o DW no diagrams.io do Google.

Cronograma do Trabalho

considerando a data final como dia 23/06 temos o seguinte cronograma.

Entrega 1 20/05	Implementação do DW na ferramenta PostgreSQL e utilizando pentaho.	Pessoas responsáveis: Lucas e Felipe
Entrega 2 26/05	Começo da escrita do relatório final no modelo artigo SBC.	Pessoas responsáveis: Lucas e Celso
Entrega 3 31/05	Implementação de algoritmos de aprendizado de máquinas para séries temporais e seus resultados(métricas) em python utilizando sklearn.	Pessoas responsáveis: Celso e Felipe
Entrega 4 05/06	Adicionar as métricas e continuação da escrita do artigo e toques finais no processo de BI.	Pessoas responsáveis: Lucas, Felipe e Celso
Entrega 5(final) 20/06	Artigo completo, algoritmos, resultados e implementações.	Pessoas responsáveis: Lucas, Felipe e Celso

Referências

https://covid.saude.gov.br/ dados covid-19 Brasil
https://www.indexmundi.com/g/r.aspx?v=2227&l=es (leitos por 1000 habt.)
https://www.mscbs.gob.es/estadEstudios/sanidadDatos/home.htm (leitos Esp)
https://data.brasil.io/dataset/covid19/_meta/list.html (São Paulo e Rio de Janeiro)
https://github.com/datadista/datasets/blob/master/COVID%2019/ (Madrid)
https://github.com/nytimes/covid-19-data (cidades EUA - NY, LA, MI)
https://github.com/pcm-dpc/COVID-19/blob/master (Roma)
https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Argentina (Buenos Aires)

Dados climáticos>
Canberra (Austrália):
http://www.bom.gov.au/climate/dwo/202001/html/IDCJDW2801.202001.shtml (Jan)
http://www.bom.gov.au/climate/dwo/202002/html/IDCJDW2801.202002.shtml (Fev)
http://www.bom.gov.au/climate/dwo/202003/html/IDCJDW2801.202003.shtml (Mar)
http://www.bom.gov.au/climate/dwo/202004/html/IDCJDW2801.202004.shtml (Abr)

Buenos Aires (Argentina):

https://www.accuweather.com/es/ar/buenos-aires/7894/march-weather/7894?year=2020&view=list (Mar)
https://www.accuweather.com/es/ar/buenos-aires/7894/april-weather/7894?year=2020&view=list (Abr)

Nova Iorque (EUA) :
https://www.accuweather.com/en/us/new-york/10007/march-weather/349727?year=2020&view=list (Mar)
https://www.accuweather.com/en/us/new-york/10007/april-weather/349727?year=2020&view=list (Abr)

Miami (EUA)
https://www.accuweather.com/en/us/miami/33128/march-weather/347936?year=2020&view=list (Mar)
https://www.accuweather.com/en/us/miami/33128/april-weather/347936?year=2020&view=list (Abr)

Los Angeles (EUA):
https://www.accuweather.com/en/us/los-angeles/90012/january-weather/347625?year=2020&view=list (Jan)
https://www.accuweather.com/en/us/los-angeles/90012/february-weather/347625?year=2020&view=list (Fev)
https://www.accuweather.com/en/us/los-angeles/90012/march-weather/347625?year=2020&view=list (Mar)
https://www.accuweather.com/en/us/los-angeles/90012/april-weather/347625?year=2020&view=list (Abr)

São Paulo (Brasil):
https://www.accuweather.com/en/br/s%C3%A3o-paulo/45881/february-weather/45881?year=2020&view=list (Fev)
https://www.accuweather.com/en/br/s%C3%A3o-paulo/45881/march-weather/45881?year=2020&view=list (Mar)
https://www.accuweather.com/en/br/s%C3%A3o-paulo/45881/april-weather/45881?year=2020&view=list (Abr)

Rio de Janeiro (Brasil):
https://www.accuweather.com/en/br/rio-de-janeiro/45449/march-weather/45449?year=2020&view=list (Mar)
https://www.accuweather.com/en/br/rio-de-janeiro/45449/april-weather/45449?year=2020&view=list (Abr)

Madrid (Espanha):
https://www.accuweather.com/en/es/madrid/308526/february-weather/308526?year=2020&view=list (Fev)

https://www.accuweather.com/en/es/madrid/308526/march-weather/308526?year=2020&view=list (Mar)
https://www.accuweather.com/en/es/madrid/308526/april-weather/308526?year=2020&view=list (Abr)

Roma (Itália):
https://www.accuweather.com/en/it/rome/2-213490_1_al/february-weather/2-213490_1_al?year=2020&view=list (Fev)
https://www.accuweather.com/en/it/rome/2-213490_1_al/march-weather/2-213490_1_al?year=2020&view=list (Mar)
https://www.accuweather.com/en/it/rome/2-213490_1_al/april-weather/2-213490_1_al?year=2020&view=list (Abr)