

UFRJ - Data Warehouse (MAB 602)

Grupo 06: Thiago Outeiro Pereira Damasceno - DRE: 116038363

Giovani Tricarico Barros - DRE: 118051317

João Wendling Assayag - DRE: 118113834

Professor: Geraldo Xexeo

1. Coleta de dados

Inicialmente nosso grupo faria o trabalho em grupo baseado no site [“Painel COVID RJ”](#)¹, porém, os dados do site se mostraram insuficientes para elaboração do trabalho, com isso, decidimos buscar novos dados, chegando a Johns Hopkins University (Universidade privado norte-americana), que disponibilizam os dados que são mostrados pelo Google, ao realizar uma busca por “COVID 19 Brasil”. Para tal, eles utilizam como fonte um [repositório do Github](#)² que mostram dados disponibilizados pela UFV (Universidade Federal de Viçosa), nesse repositório existem diversos tipos de dados, desde informações mais gerais sobre a Covid no país, a dados específicos em cada cidade. Para acompanhar a evolução do vírus, os dados mais interessantes foram os [dados por estado](#)³.

Para leitura dos dados, foi utilizado o *node* “FileReader” com a configuração de “read from” como “Custom/KNIME URL” e passando a URL dos dados, o resto foi deixado como auto detecção.

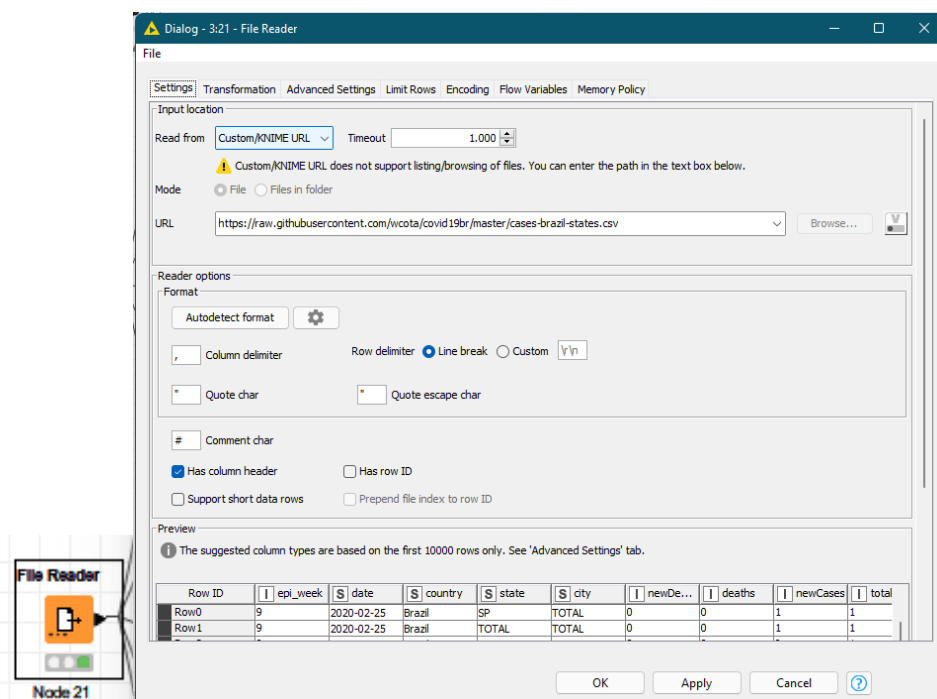


Imagem 1. Leitura de dados no Knime

2. Diagrama Estrela

Para a criação do modelo estrela, foi observado os dados presentes no arquivo, sendo esses: localidade, data, vacinação e evolução.

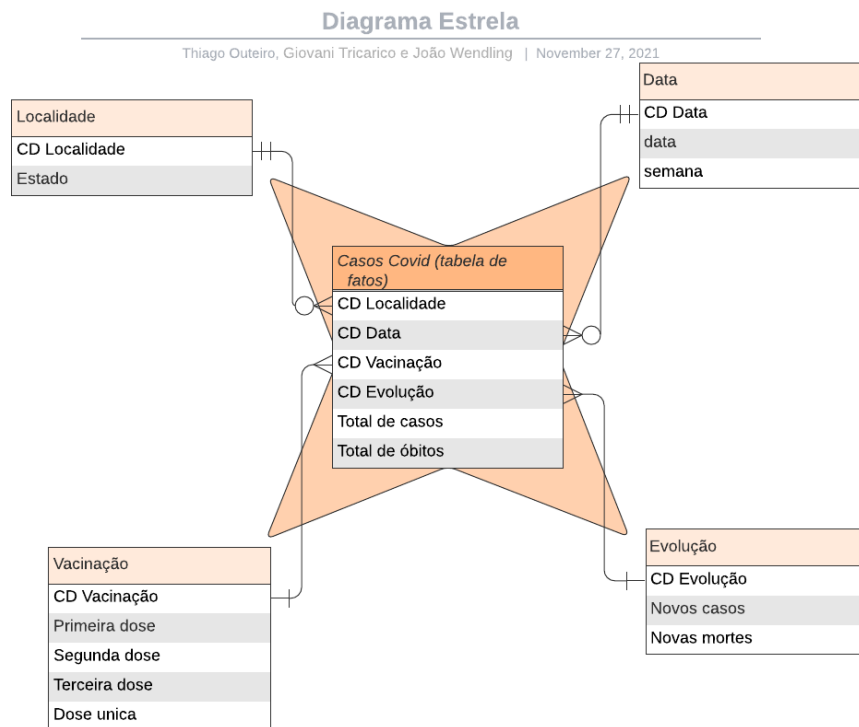


Imagem 2. Diagrama Estrela – Casos Covid Brasil

3. Criação do banco de dados

Para a criação do banco de dados, foi utilizado o aplicativo *mySQL Workbench*, e criado uma *connection* na própria IDE como mostrado abaixo:

```
1 • create schema provaGrupoDW;  
2 • use provaGrupoDW;
```

Imagem 3. Conexão feita

Para a criação de tabelas foi utilizado o KNIME, no momento da carga dos dados, utilizando os nós *MySQL Connector* e *DBWrite*:

1. **MySQL Connector:** Utilizado para se conectar com o banco criado no *MySQL WorkBench*, sua saída é o banco que irá se conectar na entrada no DB Writer.
2. **DBWriter:** Utilizado para escrever no banco de dados recebido do *MySQL Connector*, irá criar a tabela passada nas configurações caso ela ainda não exista.

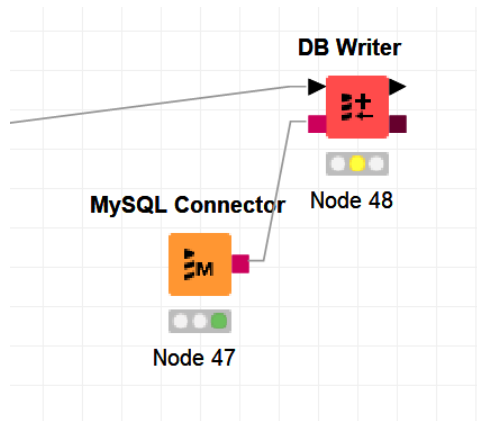


Imagem 4. Uso dos nós

Para configurar esses nós é necessário que o banco seja criado no *mySql Workbench* e o *schema* seja criado com o seguinte script:

- 1 • `create schema provaGrupoDW;`
- 2 • `use provaGrupoDW;`

Imagem 5. Configurações dos nós

Após isso no *node* de *Connector* será necessário selecionar o banco (no caso foi utilizado o próprio *main*) e o no *writer* será selecionado o *schema* e a *tabela* a ser criada.

4. Carregamento dos dados

Para o carregamento dos dados foi utilizado o banco criado na questão 3, em conjunto com o KNIME obtendo o seguinte fluxo:

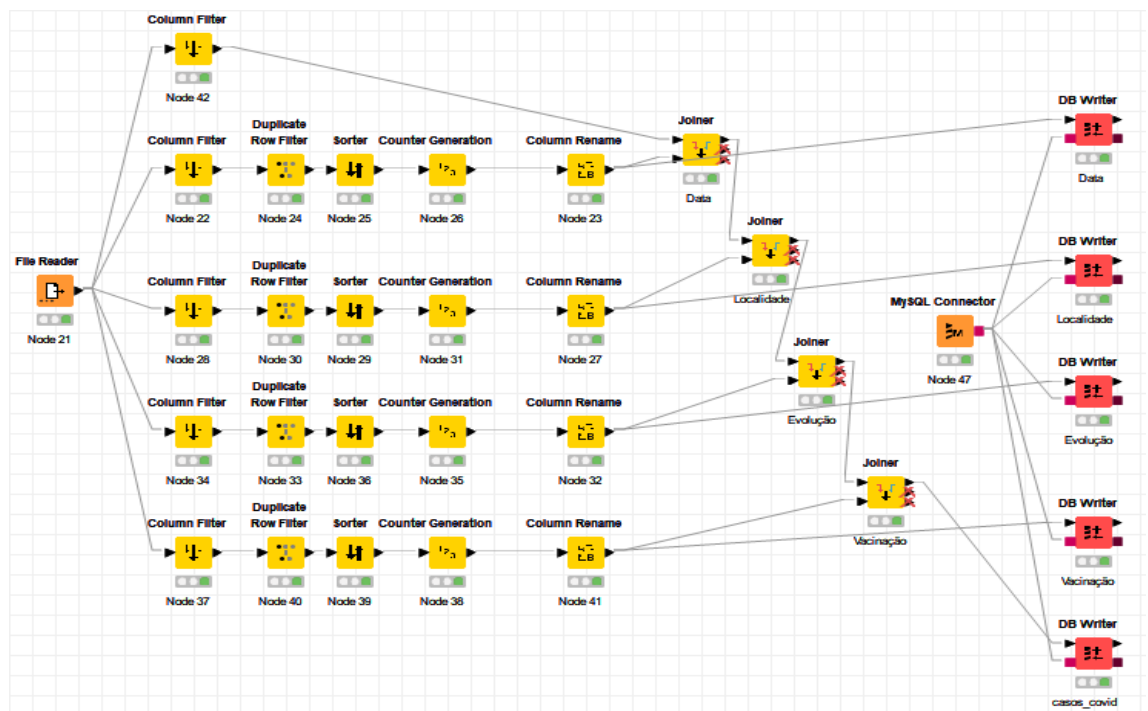


Imagem 6. Configurações dos nós

Para a carga de dados foram utilizados os *nodes*:

- **Column Filter:** Utilizado para selecionar apenas as colunas necessárias para as tabelas.
- **Duplicate Row Filter:** Utilizado nas dependências para remover as linhas duplicadas, ficando com entradas únicas.
- **Sorter:** Não é necessário, porém foi utilizado para a visualização ficar melhor, organizando as linhas das tabelas.
- **Counter Generation:** Utilizado para gerar os IDs (chaves) das tabelas de dependências
- **Column Rename:** Utilizado para renomear as tabelas com o mesmo nome especificado no modelo.

Além destes 5 nodes, também foi utilizado o Node de Joiner, para realizar o pareamento da tabela de fatos com a chave para as de dependências. Todos os Joiners foram Inner Join, menos o último onde foi realizado o left outer join, pois como se tratam de dados de vacinas, no início dos casos não haviam vacinas. Como os dados eram concisos, o número de linhas entre o início e o final foi igual (17486). Com isso na interface do banco temos:

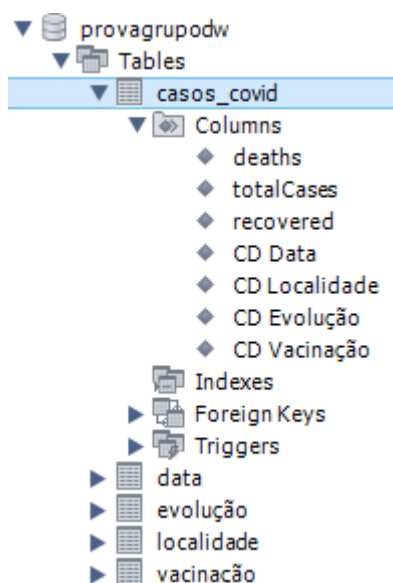






Imagem 7. Interface do banco de dados

Realizando um select na tabela de casos_covid:

```
3 • use provaGrupoDW;
4 • select * from casos_covid;
```

Result Grid									 Filter Rows:	<input type="text"/>	Export:		Wrap Cell Content:		Fetch
	deaths	totalCases	recovered	CD Data	CD Localidade	CD Evolução	CD Vacinação								
▶	9403	369475	354985	555	11	4183	4402								
	582243	20837967	19138604	555	27	13167	7717								
	9411	369733	357475	556	11	3101	4408								
	582983	20861571	19243454	556	27	13145	7718								
	9418	369916	357475	557	11	2703	4412								
	583621	20881896	19243454	557	27	13072	7719								
	9429	370014	357475	558	11	4122	4414								
	583892	20894557	19246432	558	27	12649	7720								
	12313	567037	542182	559	7	4276	5333								
	9434	370224	357475	559	11	1931	4415								
	146567	4291993	4022314	559	25	6195	7517								
	584112	20904729	19246432	559	27	12468	7721								
	12318	567686	542182	560	7	2133	5335								
	9439	370341	357475	560	11	1866	4418								
	146595	4295149	4022314	560	25	8033	7518								
	584439	20918333	19246432	560	27	12788	7722								
	12333	568392	542182	561	7	5401	5345								
	9447	370399	357475	561	11	3021	4421								
	146610	4297229	4055982	561	25	5517	7519								
	584688	20932515	19295628	561	27	12581	7723								
	12341	569268	542182	562	7	3327	5350								
	9455	370399	357475	562	11	3008	4422								
	146828	4298180	4055982	562	25	12458	7521								
	585512	20965338	19296570	562	27	13189	7724								
	12350	570074	542182	563	7	3685	5356								
	9459	370454	357475	563	11	1424	4428								
	147020	4298851	4055982	563	25	12341	7524								
	586140	20981303	19307340	563	27	13071	7725								

casos_covid2

Imagem 8. Tabela casos Covid

5. Análise de dados

Após a coleta dos dados e a implementação do modelo dimensional, foi observado e pensado as seguintes perguntas:

1. Qual foi a taxa média de contaminação e óbito pela Covid-19 no Brasil?
2. Como funciona a taxa de Redução da evolução da doença após cada etapa de implementação das doses da vacina?

Qual é a quantidade de casos e óbitos por estado desde o início da pandemia?

3. Qual foi o mês com maiores números de novos casos desde o começo da pandemia?
4. Qual foi o desenvolvimento diário de novos casos de óbitos pela Covid-19?

Com as perguntas já feitas, optamos pela de número 3 para ser respondida.

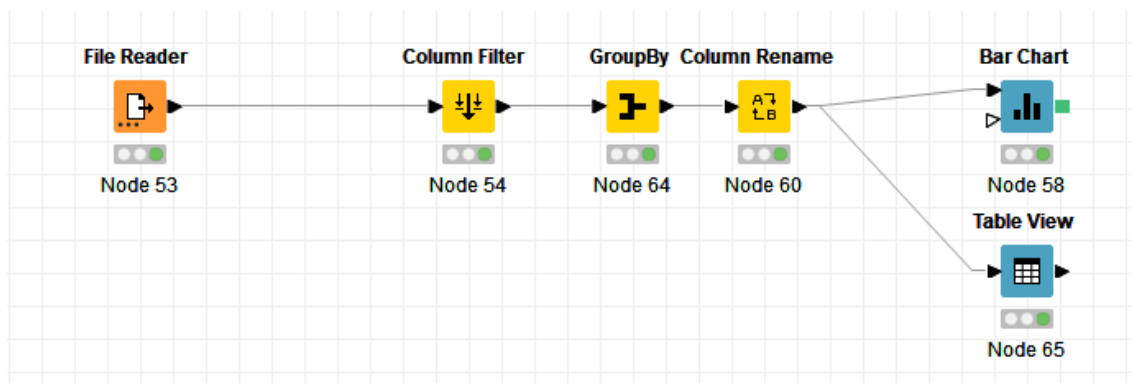


Imagem 9. Nós para leitura dos casos e óbitos por estado

Para começar a análise de dados, a partir da tabela dos casos de COVID, foi filtrado as colunas que representam a quantidade de novos casos, óbitos e estados no Brasil. As linhas representavam as atualizações diárias de cada estado, conforme disponibilizado, assim, como queríamos o total de casos por estado desde o começo da pandemia, agrupamos por localização, todas as linhas que apresentavam o valor correspondente a cada estado, somando assim todos os casos de óbito e contaminação já registrados encontrando assim a totalidade do número de casos e óbitos em cada unidade federativa como é visível nas imagens 10 e 11 abaixo.

Row ID	state	Mortes	Casos
Row0	AC	1846	88203
Row1	AL	6346	241409
Row2	AM	13801	429743
Row3	AP	2002	124528
Row4	BA	27282	1258872
Row5	CE	24624	949590
Row6	DF	11026	517608
Row7	ES	13157	619366
Row8	GO	24504	934821
Row9	MA	10281	364672
Row10	MG	56143	2206864
Row11	MS	9682	378510
Row12	MT	13997	552439
Row13	PA	16886	607981
Row14	PB	9520	460269
Row15	PE	20226	639373
Row16	PI	7185	331273
Row17	PR	40771	1577632
Row18	RJ	69011	1342671
Row19	RN	7486	381183
Row20	RO	6633	276791
Row21	RR	2050	128348
Row22	RS	36075	1489564
Row23	SC	19973	1231919
Row24	SE	6043	278691
Row25	SP	153993	4437386
Row26	TO	3916	232335
Row27	TOTAL	614459	22082041

Imagem 10. Tabela de casos e óbitos por estado

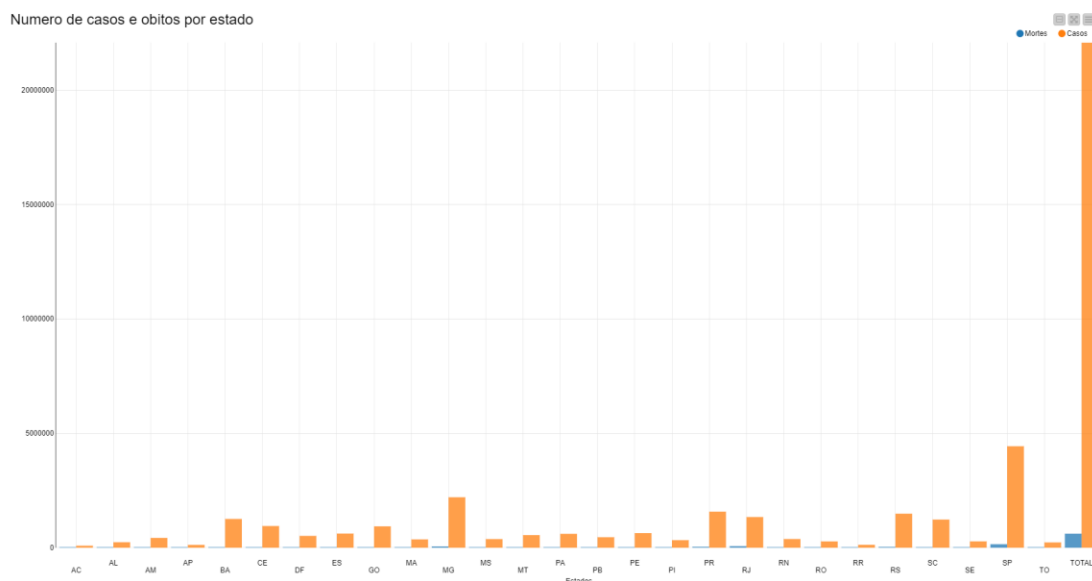


Imagem 11. Gráfico de casos e óbito por estado

Observando esses dados, é notório que existe uma disparidade na quantidade de casos de contaminação e óbitos (principalmente na região Sudeste, onde estão concentrados a maior parte dos casos). Foi constatada também uma diferença notória na taxa de mortalidade e de contaminação, quando comparado aos números registrados por cada estado, podendo concluir que é uma doença extremamente propagativa.

6. Aprendizado

Para o aprendizado foi utilizado o KNIME, com a seguinte pergunta: “É possível saber qual o estado com base no número de casos total e a semana?”, para isso, a tabela foi filtrada, agrupada por casos totais e semanas, particionada em 70% para o aprendizado e 30% para a predição, seguindo o seguinte fluxo:

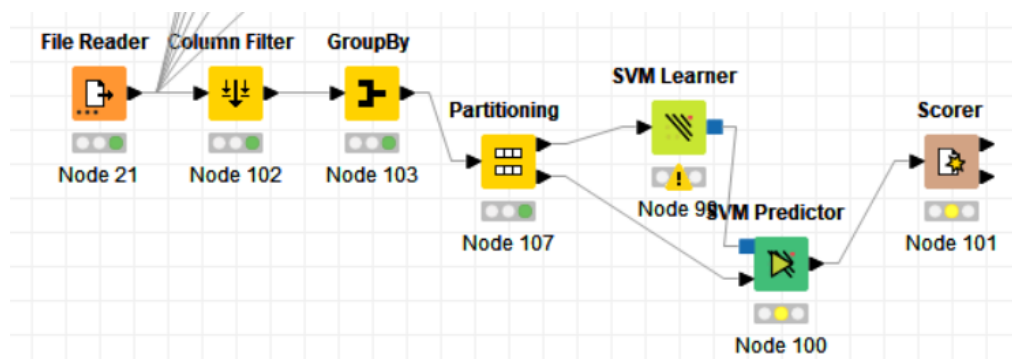


Imagem 12. Gráfico de casos e óbito por estado

No nó de aprendizado foi selecionada a coluna “State” para ser aprendida e o método polinomial, porém ao tentar rodar, o número máximo de interações era alcançado, mesmo havendo um pequeno número de linhas, o que fazia com que a predição não funcionasse corretamente.

7. Referencia

1. <https://experience.arcgis.com/experience/38efc69787a346959c931568bd9e2cc4>
2. <https://github.com/wcota/covid19br>
3. <https://raw.githubusercontent.com/wcota/covid19br/master/cases-brazil-states.csv>

Nosso repositório no Git: https://github.com/topd97/prova_grupo_dw_2021/