



Pós Graduação em Ciência de Dados e Analytics

MVP – Engenharia de dados

Aluno: Clarisse L Sieczko

1. Objetivo

O objetivo deste trabalho é avaliar qual o impacto que a pandemia teve no ENEM.

Questões a serem respondidas:

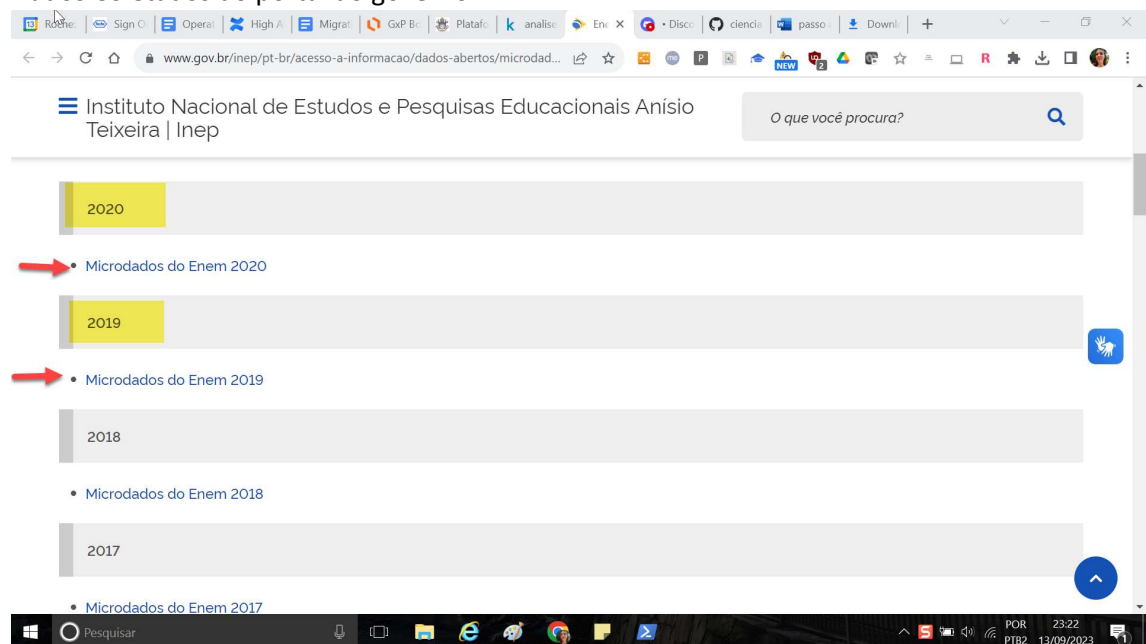
- Estudantes tiveram seu desempenho reduzido no enem após a pandemia?
- Há diferença/queda de rendimento entre alunos de escolas públicas e privadas comparando antes e depois da pandemia?

2. Detalhamento

a. Busca pelos dados

A realização de busca pelos dados deu-se utilizando o portal do governo <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem> Foram baixados dados dos anos de 2019 (antes da pandemia) e 2020 (1 ano após início da pandemia) para a máquina e depois feito upload dos arquivos em formato .CSV no Blob storage do Azure (clarissemvp)

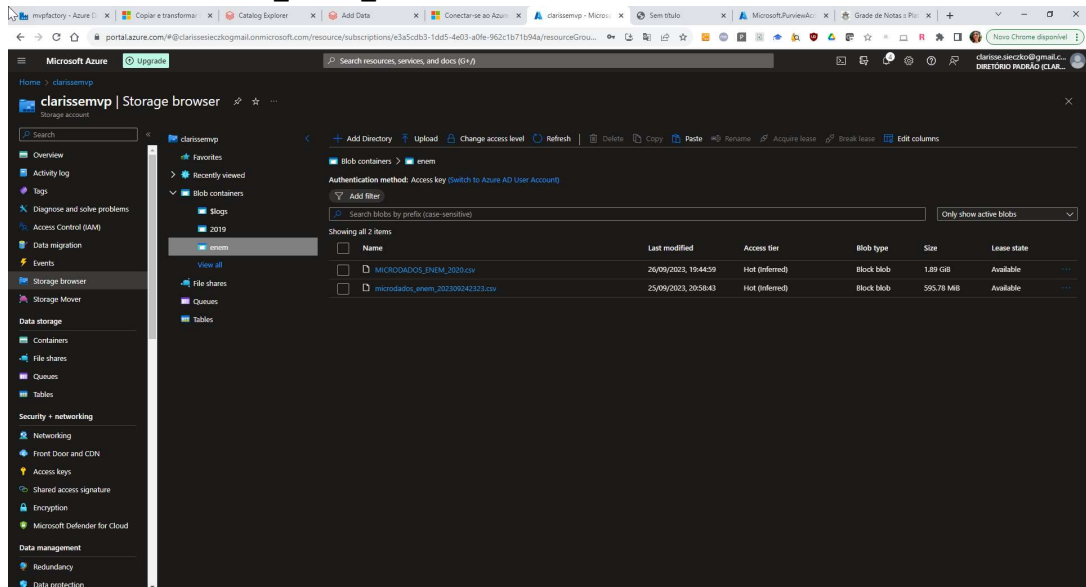
Dados Coletados do portal do governo:



b. Carga inicial

Dados no Azure Blob Storage:

- microdados_enem_202309242323.csv – Dados de 2019
- MICRODADOS_ENEM_2020.csv – Dados de 2020

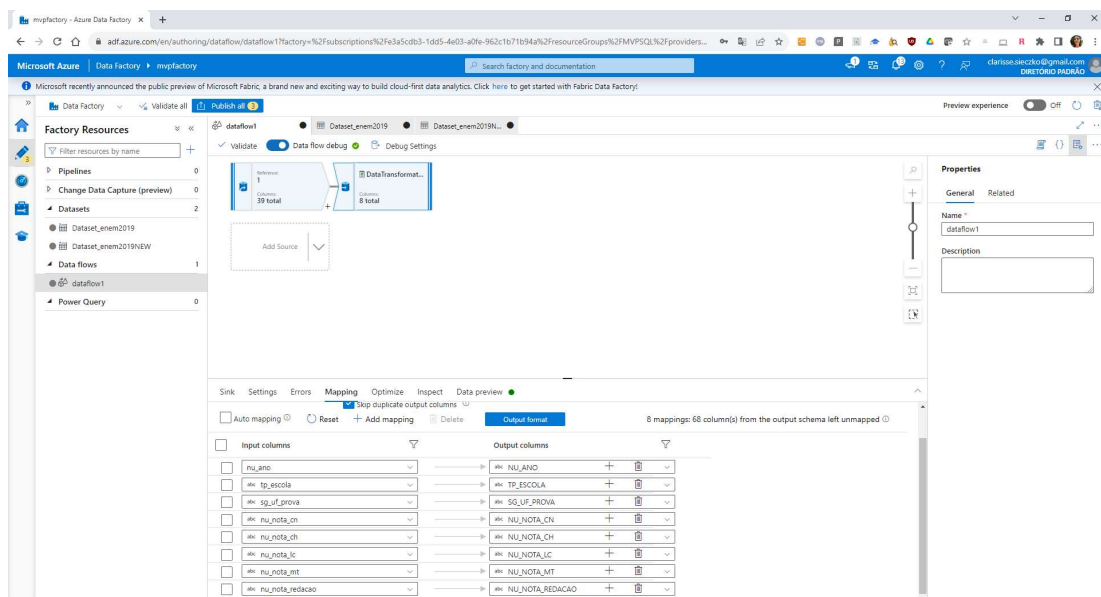


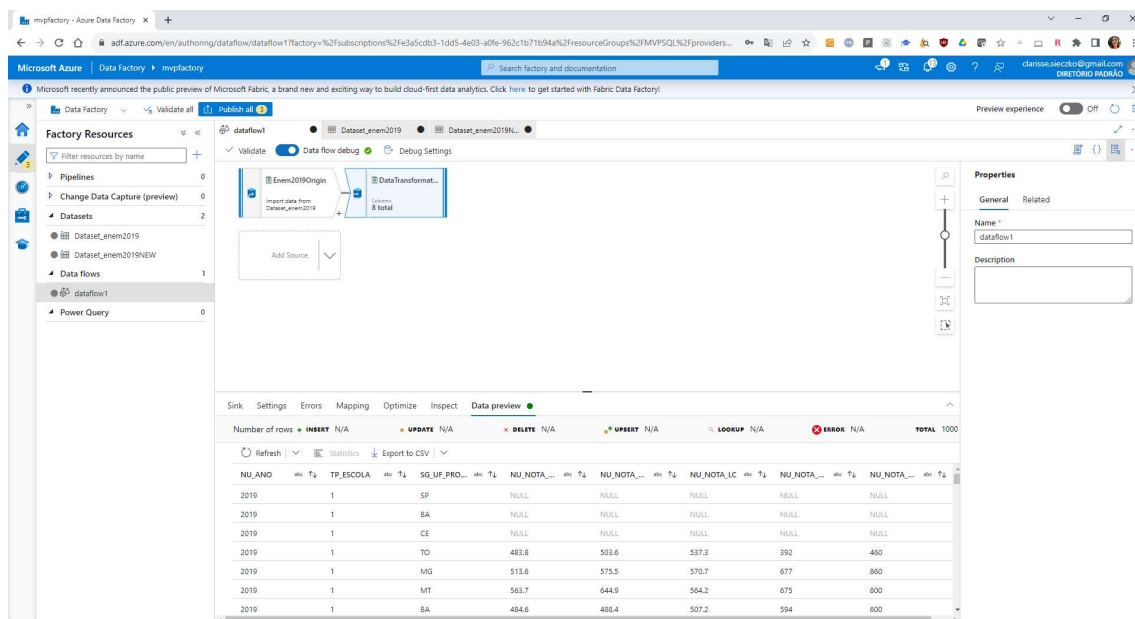
c. Modelagem

A modelagem foi realizada nos arquivos .CSV, armazenado no Azure Bloob Storage, utilizando a ferramenta ETL Azure Data Factory.

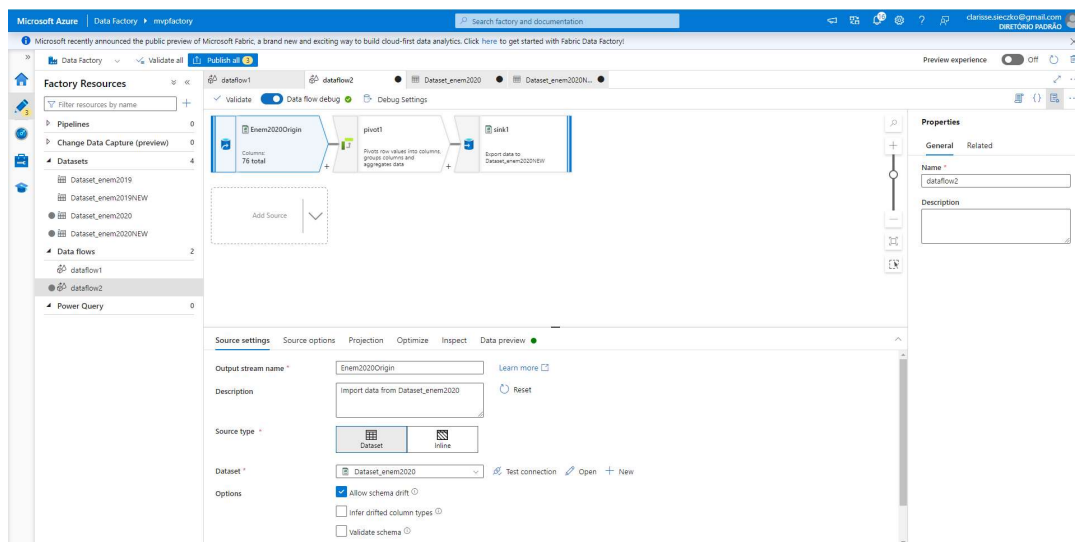
Foram criados Dataflows que executaram as seguintes transformações:

1) Arquivo microdados_enem_202309242323.csv de 2019, continha 39 colunas e foi usado o Sink para mapear e reduzir para um arquivo .CSV com 8 colunas que serão necessárias para realizar as análises.





2) Arquivo MICRODADOS_ENEM_2020.csv de 2020, foi utilizado o pivot para transformar texto delimitado em colunas e também o Sink para mapear e reduzir a um arquivo .CSV com 8 colunas que serão necessárias para realizar as análises.



Após esse passo, um Pipeline foi criado para realizar a cópia dos dados em formato .csv para um banco de dados Azure SQL.

Microsoft Azure | Data Factory | myfactory

Factory Resources

- Pipelines: 1
 - data
- Datasets: 6
 - AzureSqlTable1
 - AzureSqlTable2
 - Dataset_enem2019
 - Dataset_enem2019NEW
 - Dataset_enem2020
 - Dataset_enem2020NEW
- Data flows: 2
 - dataflow1
 - dataflow2
- Power Query: 0

Activities

- Move and transform
 - Copy data
 - Copy data1
 - Copy data1_copy1
- Data flow
- Azure Data Explorer
 - Azure Data Explorer C...
- Databricks
- Notebook
- Jar
- Python
- Data Lake Analytics
 - U-SQL
- General
 - Get Metadata

Pipeline run ID: 7b4be660-a3b8-468c-9197-aa6b5a70c3ea

Pipeline status: Succeeded

Activity run ID: a5f8aa8-0643-4d85-8c58-af80ecd87331

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Copy data1_copy1	Succeeded	Copy data	9/28/2023, 6:22:09 PM	8m 17s	AutoResolveIntegration	a5f8aa8-0643-4d85-8c58-af80ecd87331	
Copy data1	Succeeded	Copy data	9/28/2023, 6:22:09 PM	8m 21s	AutoResolveIntegration	1a76943e-f73b-42a3-912d-5b0ecd96c04a	

Details

Learn more on copy performance details from here.

Activity run ID: a5f8aa8-0643-4d85-8c58-af80ecd87331

Copy data1

Data read: 2.025 GB

Data written: 199.852 MB

Files read: 1

Rows read: 5,783,109

Rows written: 5,783,109

Peak connections: 11

Copy duration: 00:08:13

Throughput: 4,174 MB/s

Start time: 9/28/2023, 6:22:09 PM

Used DUs: 4

Used parallel copies: 1

Duration: 00:08:13

Details

Queue	Working duration	Total duration
Queue	00:00:07	00:00:00
Pre copy script	00:00:00	00:00:00
Transfer	00:00:00	00:00:05

Data consistency verification: Not verified

How satisfied or dissatisfied are you with the performance of this copy activity?

★★★★★

Details

Learn more on copy performance details from here.

Activity run ID: 1a76943e-f73b-42a3-912d-5b0ecd96c04a

Copy data1

Data read: 634.718 MB

Data written: 232.08 MB

Files read: 1

Rows read: 5,095,171

Rows written: 5,095,171

Peak connections: 11

Copy duration: 00:08:19

Throughput: 1,272 MB/s

Start time: 9/28/2023, 6:22:09 PM

Used DUs: 4

Used parallel copies: 1

Duration: 00:08:19

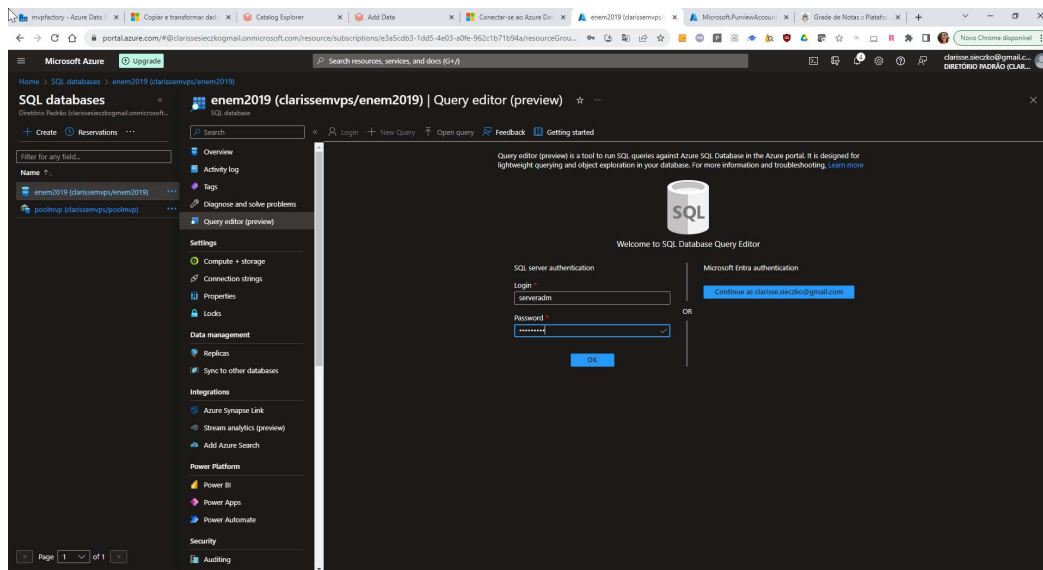
Details

Queue	Working duration	Total duration
Queue	00:00:06	00:00:00
Pre copy script	00:00:00	00:00:00
Transfer	00:00:00	00:00:11

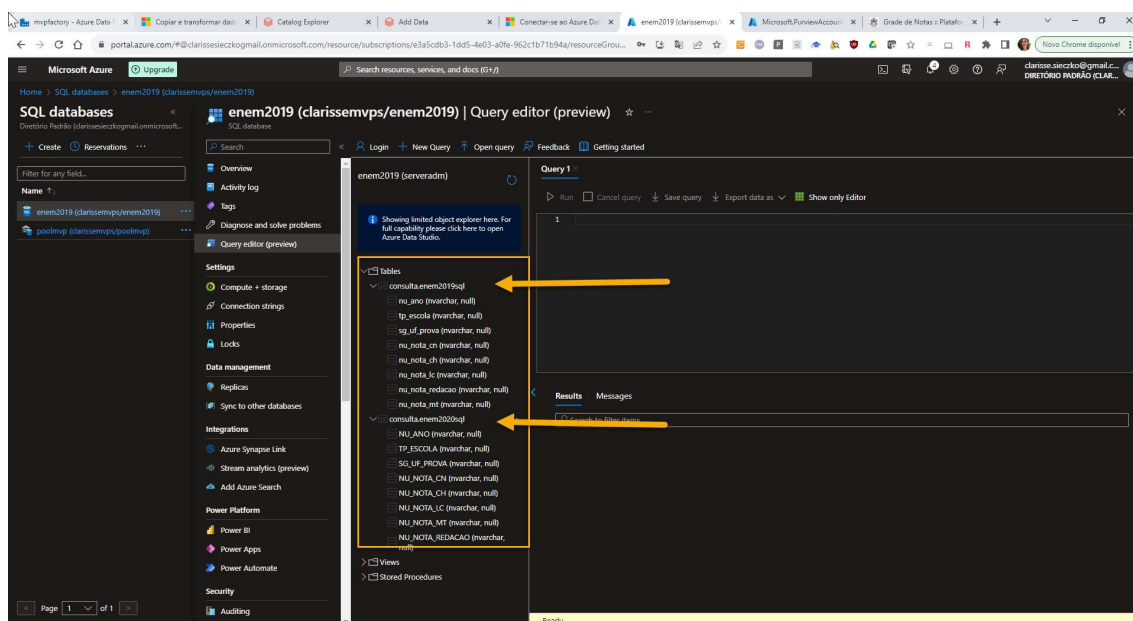
Data consistency verification: Not verified

How satisfied or dissatisfied are you with the performance of this copy activity?

★★★★★



Foi criado uma instância e dentro do database, encontram-se as duas tabelas criadas a partir dos arquivos .csv



Catálogo de dados

Não foi possível criar um catálogo de dados na cloud da Azure, pois o tipo de assinatura não permite. Outra tentativa foi utilizar o Pureview, mas não consegui realizar a conexão ao banco de dados.

Portanto, o catálogo será descrito manualmente aqui neste documento.

Microsoft Azure | DIRETÓRIO PADRÃO

Home > Create a resource > Marketplace >

New data catalog

Provide data catalog info

Basics Tags Review + Create

Create a Data catalog to manage and contain all of your data entities. [Learn more.](#)

Project details

Subscription * Azure subscription 1

Resource group * clarisse_MVP [Create new](#)

Instance details

Data catalog name * Datalake

✖ You've logged in using a personal Microsoft account. Azure Data Catalog only supports using work or school accounts. Please log in using your work or school account to create an Azure Data Catalog.

Location * East US

Pricing tier * Free

[Review + Create](#) [Previous](#) [Next: Tags >](#)

Nome do Banco de Dados	Descrição	Localização Física	Tamanho do Banco de Dados
Consulta.enem2019sql	Banco de dados com resultados das provas do Enem do ano de 2019	Cloud Azure	232,08 MB
Consulta.enem2020sql	Banco de dados com resultados das provas do Enem do ano de 2020	Cloud Azure	199,852 MB

Atributo da tabela	Tipo	Descrição
NU_ANO	varchar	Ano da prova
TP_ESCOLA	varchar	Tipo de Escola (1- Federal, 2- Estadual, 3- Municipal, 4- Particular)
SG_UF_PROVA	varchar	Local da prova (Unidades Federais do Brasil: RJ, BA, etc)
NU_NOTA_CN	varchar	Nota bruta prova de Ciências da Natureza
NU_NOTA_CH	varchar	Nota bruta prova de Ciências Humanas
NU_NOTA_LC	varchar	Nota bruta prova de Linguagens e Código
NU_NOTA_MT	varchar	Nota bruta prova de Matemática
NU_NOTA_REDACAO	varchar	Nota bruta prova de redação

d. Análise

Qualidade de dados

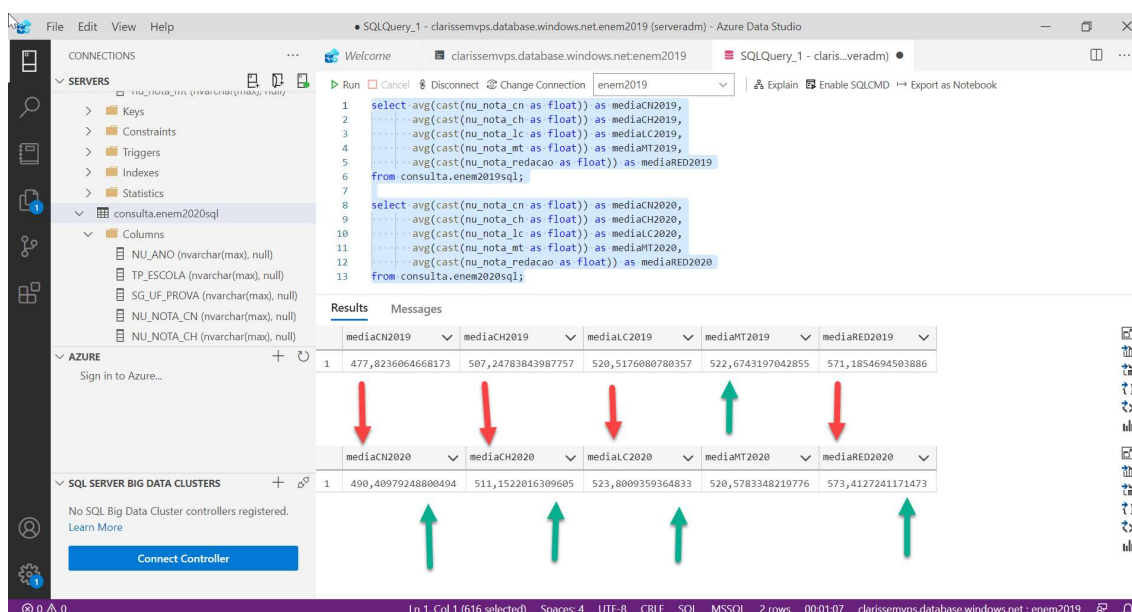
Os dados já encontravam-se curados, não houve necessidade de realizar nenhum tratamento em relação à qualidade dos dados.

Porém, durante a análise percebi que haviam muitos campos nulos. Mas ao invés de tratar esses campos, decidi que o fato deles serem nulos me traria mais uma informação que não havia considerado no início da descrição do problema, que é o quanto a pandemia afetou a execução da prova do Enem em relação ao absenteísmo.

Solução do problema

- Estudantes tiveram seu desempenho reduzido no enem após a pandemia?

Foi realizado o comparativo das médias de notas de cada uma das disciplinas do Enem em relação a 2019 (antes da pandemia) e 2020 (1 ano de pandemia).



Como resultado da consulta, olhando basicamente para o quesito nota, pode-se ver que não houve queda no desempenho dos alunos no Enem por causa da pandemia. A maioria das provas teve desempenho ligeiramente superior no ano de 2020, inclusive. O resultado foi uma grande surpresa, pois a pandemia da COVID afetou os estudos drasticamente, pois assim como muitas outras áreas, as escolas ficaram um longo período sem aulas e após algum tempo, algumas escolas retornaram os estudos online.

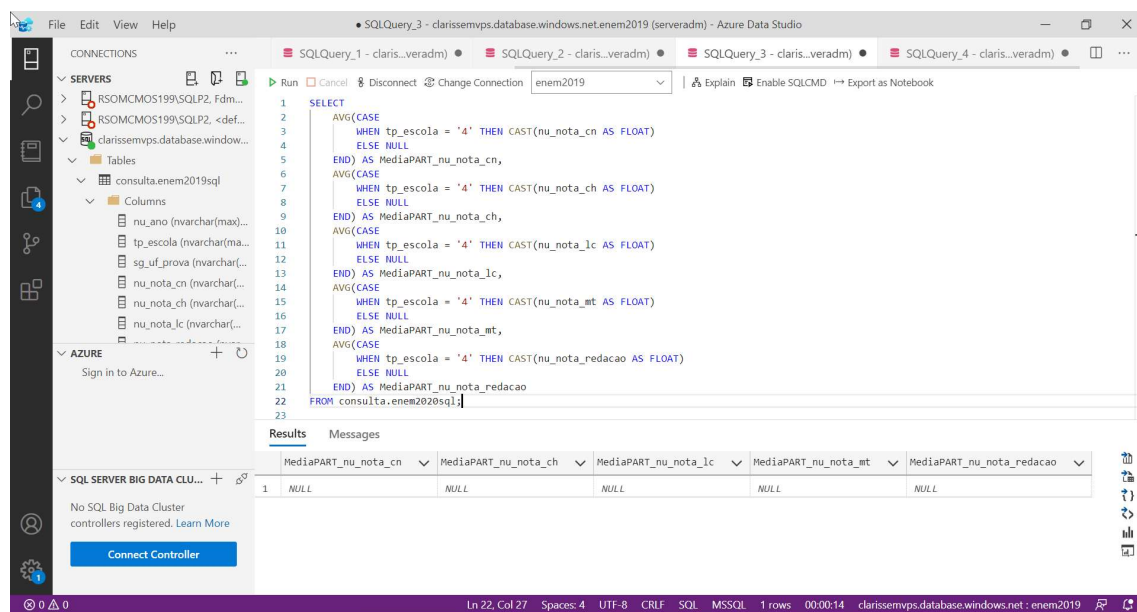
As perguntas seguintes, ajudarão a compor a análise e detectar afinal, se e como a pandemia afetou os estudantes que prestaram prova do Enem.

- Há diferença/queda de rendimento entre alunos de escolas públicas e privadas comparando antes e depois da pandemia?

O intuito para responder essa pergunta seria consultar as médias por tipo de escola: 1, 2 e 3 pública, 4 particular. E comparar entre os anos de 2019 e 2020.

Mas ao realizar a consulta, não retornou escolas particulares em nenhuma das tabelas (consulta_enem2019sql e consulta_enem2020sql).

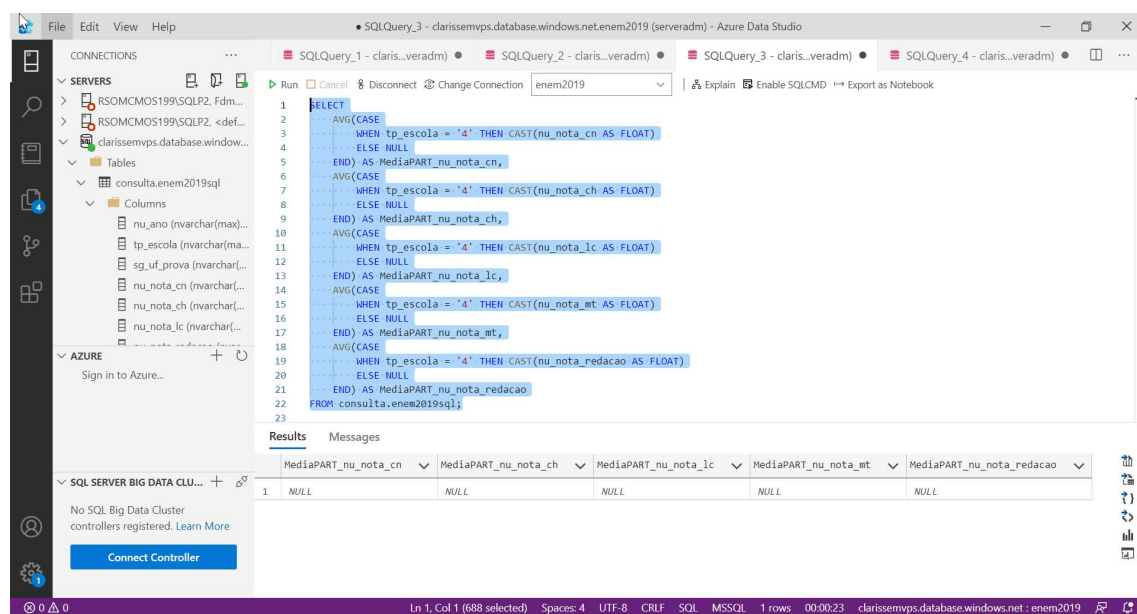
=> 2020:



```
SELECT
1
2  AVG(CASE
3    WHEN tp_escola = '4' THEN CAST(nu_notas_cn AS FLOAT)
4    ELSE NULL
5  END) AS MediaPART_nu_notas_cn,
6  AVG(CASE
7    WHEN tp_escola = '4' THEN CAST(nu_notas_ch AS FLOAT)
8    ELSE NULL
9  END) AS MediaPART_nu_notas_ch,
10 AVG(CASE
11    WHEN tp_escola = '4' THEN CAST(nu_notas_lc AS FLOAT)
12    ELSE NULL
13 END) AS MediaPART_nu_notas_lc,
14 AVG(CASE
15    WHEN tp_escola = '4' THEN CAST(nu_notas_mt AS FLOAT)
16    ELSE NULL
17 END) AS MediaPART_nu_notas_mt,
18 AVG(CASE
19    WHEN tp_escola = '4' THEN CAST(nu_notas_redacao AS FLOAT)
20    ELSE NULL
21 END) AS MediaPART_nu_notas_redacao
22 FROM consulta_enem2020sql;
```

MediaPART_nu_notas_cn	MediaPART_nu_notas_ch	MediaPART_nu_notas_lc	MediaPART_nu_notas_mt	MediaPART_nu_notas_redacao
1	NULL	NULL	NULL	NULL

=> 2019:



```
SELECT
1
2  AVG(CASE
3    WHEN tp_escola = '4' THEN CAST(nu_notas_cn AS FLOAT)
4    ELSE NULL
5  END) AS MediaPART_nu_notas_cn,
6  AVG(CASE
7    WHEN tp_escola = '4' THEN CAST(nu_notas_ch AS FLOAT)
8    ELSE NULL
9  END) AS MediaPART_nu_notas_ch,
10 AVG(CASE
11    WHEN tp_escola = '4' THEN CAST(nu_notas_lc AS FLOAT)
12    ELSE NULL
13 END) AS MediaPART_nu_notas_lc,
14 AVG(CASE
15    WHEN tp_escola = '4' THEN CAST(nu_notas_mt AS FLOAT)
16    ELSE NULL
17 END) AS MediaPART_nu_notas_mt,
18 AVG(CASE
19    WHEN tp_escola = '4' THEN CAST(nu_notas_redacao AS FLOAT)
20    ELSE NULL
21 END) AS MediaPART_nu_notas_redacao
22 FROM consulta_enem2019sql;
```

MediaPART_nu_notas_cn	MediaPART_nu_notas_ch	MediaPART_nu_notas_lc	MediaPART_nu_notas_mt	MediaPART_nu_notas_redacao
1	NULL	NULL	NULL	NULL

Apenas para as escolas públicas haviam dados reportados

```
SELECT
  AVG(CASE
    WHEN tp_escola = '1' THEN CAST(nu_nota_cn AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_cn,
  AVG(CASE
    WHEN tp_escola = '1' THEN CAST(nu_nota_ch AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_ch,
  AVG(CASE
    WHEN tp_escola = '1' THEN CAST(nu_nota_ic AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_ic,
  AVG(CASE
    WHEN tp_escola = '1' THEN CAST(nu_nota_mt AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_mt,
  AVG(CASE
    WHEN tp_escola = '1' THEN CAST(nu_nota_redacao AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_redacao
FROM consulta.enem2019sql;

SELECT
  AVG(CASE
    WHEN tp_escola = '2' THEN CAST(nu_nota_cn AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_cn,
  AVG(CASE
    WHEN tp_escola = '2' THEN CAST(nu_nota_ch AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_ch,
  AVG(CASE
    WHEN tp_escola = '2' THEN CAST(nu_nota_ic AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_ic,
  AVG(CASE
    WHEN tp_escola = '2' THEN CAST(nu_nota_mt AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_mt,
  AVG(CASE
    WHEN tp_escola = '2' THEN CAST(nu_nota_redacao AS FLOAT)
    ELSE NULL
  END) AS MediaPART_nu_nota_redacao
FROM consulta.enem2020sql;
```

MediaPART_nu_nota_cn	MediaPART_nu_nota_ch	MediaPART_nu_nota_ic	MediaPART_nu_nota_mt	MediaPART_nu_nota_redacao
479,5634687045377	509,136588915242	522,208099264446	522,296739524349	569,758188958735
469,6887837899185	490,613428594473	507,446878891354	503,76248189243765	544,9727580792916
549,5028225416593	567,7789329964676	565,608962484059	619,3180895091821	722,8381723979885

Portanto, não foi possível responder a esta pergunta.

- **Pergunta extra** (não prevista na fase inicial de definição do problema)

Antes de concluir, resolvi analisar um outro dado que só apareceu a possibilidade enquanto realizava a avaliação da qualidade dos dados. Notei que haviam valores nulos para as notas dos alunos. Então decidi analisar também se o absentismo cresceu com a pandemia.

E como resultado, podemos ver que sim, aumentou 23% em 2020 com a pandemia em relação a 2019.

```
select COUNT(*) as Absenteismo2019
from consulta.enem2019sql
where nu_nota_ch is NULL
AND nu_nota_cn is NULL
AND nu_nota_ic is NULL
AND nu_nota_mt is NULL
AND nu_nota_redacao is NULL;

select COUNT(*) as Absenteismo2020
from consulta.enem2020sql
where nu_nota_ch is NULL
AND nu_nota_cn is NULL
AND nu_nota_ic is NULL
AND nu_nota_mt is NULL
AND nu_nota_redacao is NULL;
```

Absenteismo2019
2327400

Absenteismo2020
3020210

Conclusão da análise:

Como conclusão geral da análise ao problema apresentado inicialmente, em relação aos impactos que a pandemia teve no Enem, podemos afirmar que não houve queda geral no desempenho dos alunos em relação às notas das matérias avaliadas (Ciências da Natureza, Ciências Humanas, Linguagens e Código, Matemática e Redação), sendo que inclusive houve um aumento da média de notas de 2020 em várias matérias, como: Ciências da Natureza, Ciências Humanas, Linguagens e Código e Redação.

Infelizmente como as tabelas não continham informações de escolas particulares, apesar do dicionário de dados constar, não foi possível traçar um comparativo e concluir se a pandemia afetou mais alunos da rede pública ou privada.

Porém, podemos concluir com a pergunta extra que surgiu durante as análises, que a pandemia teve um efeito negativo em relação ao absenteísmo, que cresceu 23% em 2020 (um ano de pandemia) em relação a 2019 (antes da pandemia).

Apesar de não ter conseguido responder a uma das perguntas previstas, considero que com o exposto acima, o objetivo de avaliar o impacto da pandemia no Enem foi parcialmente alcançado.

3. Autoavaliação

Considero que apesar de algumas dificuldades, foi possível atingir parcialmente o objetivo delineado no início deste trabalho.

Como dificuldades posso citar a falta de conhecimento em plataformas Cloud, ferramentas de ETL e analíticas, pois não faz parte do meu escopo de trabalho e estudos prévios.

Importante registrar minha surpresa com o resultado das médias de notas não ser negativo. Pois quando iniciei, tinha em mente um resultado muito diferente, pensei que teria defasagem no desempenho de 2020 em relação a 2019. Pensei nas dificuldades que tivemos ao longo de 2020 e como os estudantes ficaram impactados sem aulas e posteriormente online (apenas para aqueles que as escolas disponibilizaram recursos ou que os mesmos possuíam recursos como internet e computador).