Descrição

Neste trabalho, será construída um pipeline de dados utilizando tecnologias na nuvem. O pipeline irá envolver a busca, coleta, modelagem, carga e análise dos dados.
O Google Colab foi escolhida como ferramenta pois, além de ter sido abordado na sprint, é considerado como ferramenta com maior usabilidade para quem está iniciando no aprendizado de pipelines.

Durante o processo, foram abordadas as ferramentas DataFlow e DataFusion, mas foram encontrados muitos entraves no uso.

Objetivo

A partir de dados sobre as viagens pelos funcionários do serviço público, decidiu-se analisar:

- Qual órgão está com mais viagens?
- Quais funcionários estão com mais viagens?
- Quais cargos estão com mais viagens?

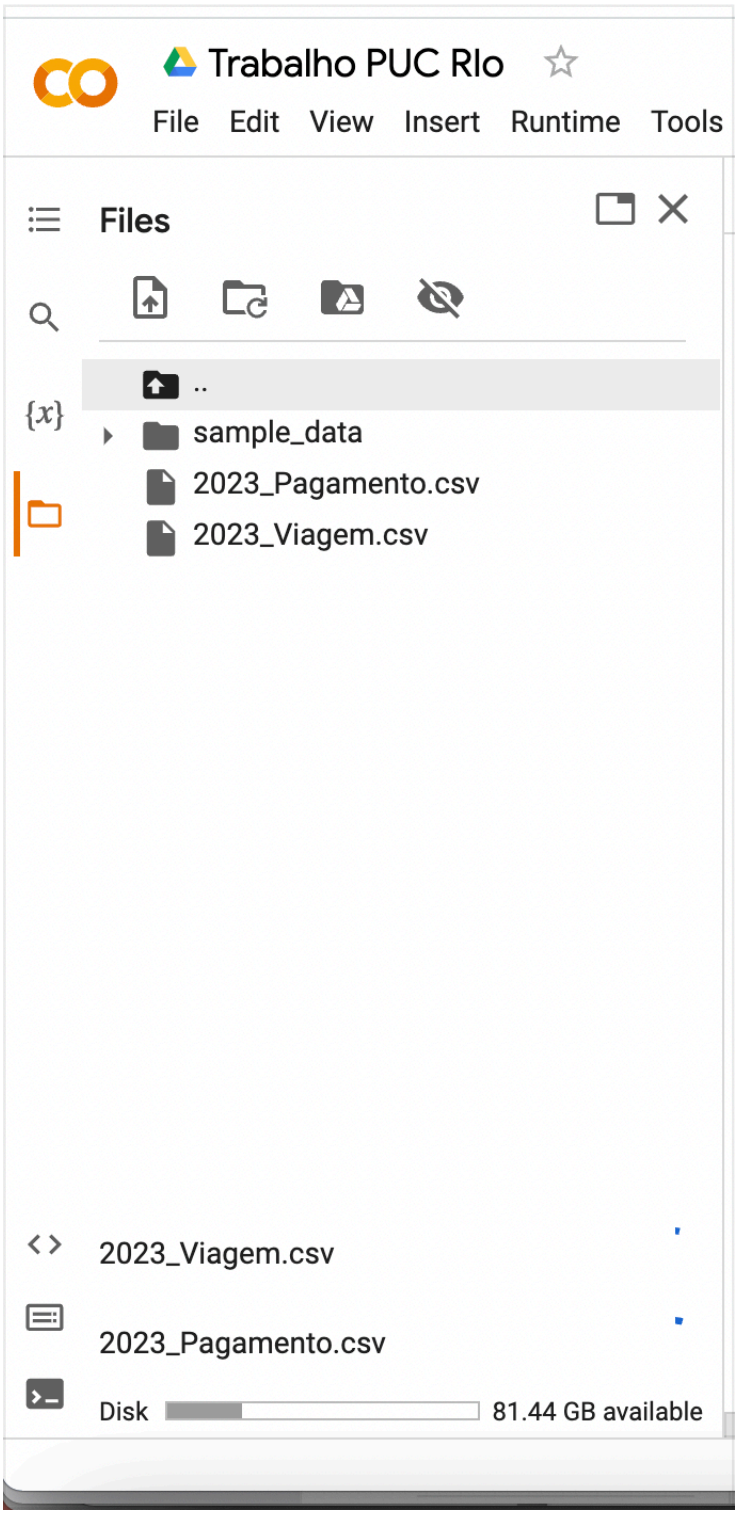
Detalhamento
Busca pelos dados

A coleta de dados foi realizada por meio do Portal da Transparência, do Governo Federal. Foram baixados os arquivos .csv Viagem e Pagamento, no intuito de responder as questões estabelecidas no objetivo desse trabalho.

https://portaldatransparencia.gov.br/download-de-dados/viagens

Coleta

Dados baixados para a máquina local e inseridos manualmente em um Bucket do Google Collab.



MODELAGEM

Catálogo de Dados

Viagem

COLUNA	DESCRIÇÃO
Identificador do processo de viagem	Número que identifica o processo de concessão da viagem

Situação	Situação da viagem: ´não realizada ´; ´realizada'
Código do Órgão Superior	Código do Órgão Superior que custeou despesas da viagem ÓRGÃO SUPERIOR - Unidade da Administração Direta que tenha entidades por ele supervisionadas. *Fonte: Manual do SIAFI*
Nome Órgão Superior	Nome do Órgão Superior
Código Solicitante	Código do Órgão que solicitou a viagem ÓRGÃO SUBORDINADO - Entidade supervisionada por um Órgão da Administração Direta. *Fonte: Manual do SIAFI*
Nome Órgão Solicitante	Nome do Órgão
CPF viajante	CPF da pessoa que realizou a viagem
Nome	Nome do viajante
Cargo	Cargo do viajante
Período - Data de início	Data de início de afastamento do servidor
Período - Data de fim	Data de fim de afastamento do servidor
Destinos	Locais pelos quais o viajante passará durante a viagem
Motivo	Motivo da viagem
Valor Diárias	Valor de diárias pagas pelo órgão, se houver
Valor Passagens	Valor de passagens pagas pelo órgão, se houver
Valor Outros Gastos	Valor de outros gastos pagos pelo órgão, se houver

Pagamento

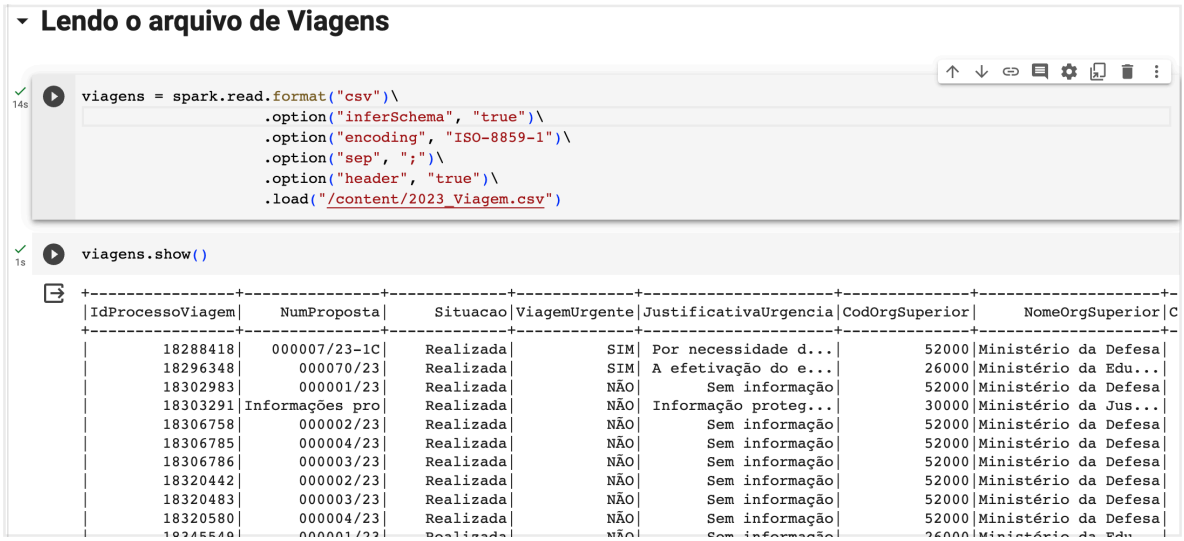
COLUNA	DESCRIÇÃO
Identificador do processo de viagem	Número que identifica o processo de concessão da viagem

Código do Órgão Superior	Código do Órgão Superior que custeou a despesa ÓRGÃO SUPERIOR - Unidade da Administração Direta que tenha entidades por ele supervisionadas. *Fonte: Manual do SIAFI*
Nome Órgão Superior	Nome do Órgão Superior
Código Órgão Pagador	Código do Órgão que pagou a despesa ÓRGÃO SUBORDINADO - Entidade supervisionada por um Órgão da Administração Direta. *Fonte: Manual do SIAFI*
Nome Órgão Pagador	Nome do Órgão
Código UG Pagadora	Código da Unidade Gestora que pagou a despesa UNIDADE GESTORA (UG) - Unidade Orçamentária ou Administrativa que realiza atos de gestão orçamentária, financeira e/ou patrimonial, cujo titular, em consequência, está sujeito a tomada de contas anual na conformidade do disposto nos artigos 81 e 82 do Decreto-lei nr. 200, de 25 de fevereiro de 1967. *Fonte: Manual do SIAFI*
Nome UG Pagadora	Nome da Unidade Gestora
Tipo de Pagamento	Tipo da despesa paga pelo órgão (diária, passagem, seguro, etc.)
Valor da Despesa	Valor da despesa paga

Carga

O ETL foi feito por meio do Google Colab, utilizando pyspark a fim de realizar manipulações dos dados para realizar a Extração, Transformação e Carga. Como resultado foi carregada uma tabela flat, contendo os dados de Viagem e Pagamento dos servidores públicos federais do ano de 2023 até o mês de setembro.

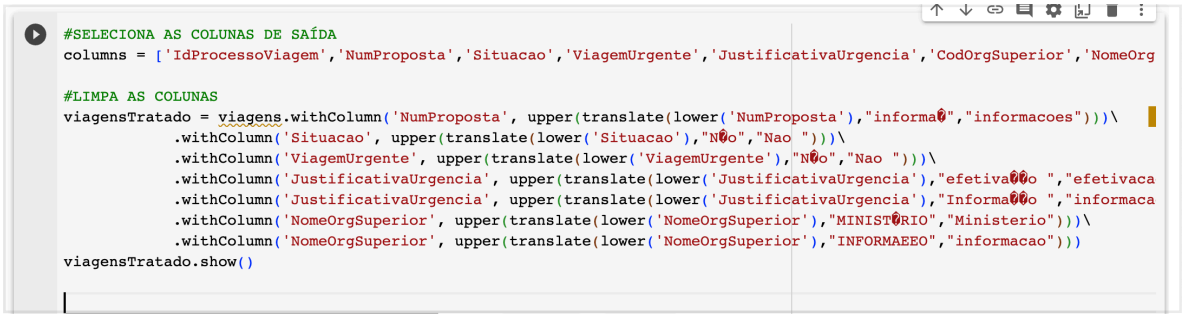
A Carga foi realizada utilizando comandos pyspark



Transformação

Passo 1 - Caracteres especiais

No primeiro momento de uso, não estava utilizando o engode correto. Então tive muitos problemas com a qualidade dos dados que estavam sendo apresentados. No intuito de tratar adequadamente a base realizei algumas instruções para corrigir caracteres especiais tais como:



No entanto, descobri o engode correto, o que possibilitou a carga dos dados de forma adequada, não precisando mais dos tratamentos acima realizados.

Passo 2 - Verificar duplicados

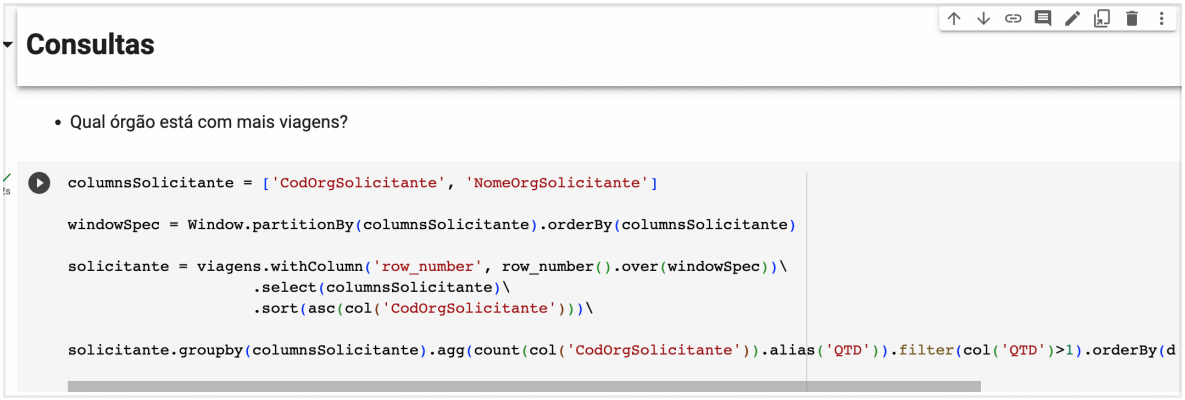
O próximo passo foi verificar se a base possuía registros duplicados.

No entanto, como pode ser evidenciado acima, a base não possuía nenhum problema dessa natureza.

Análise

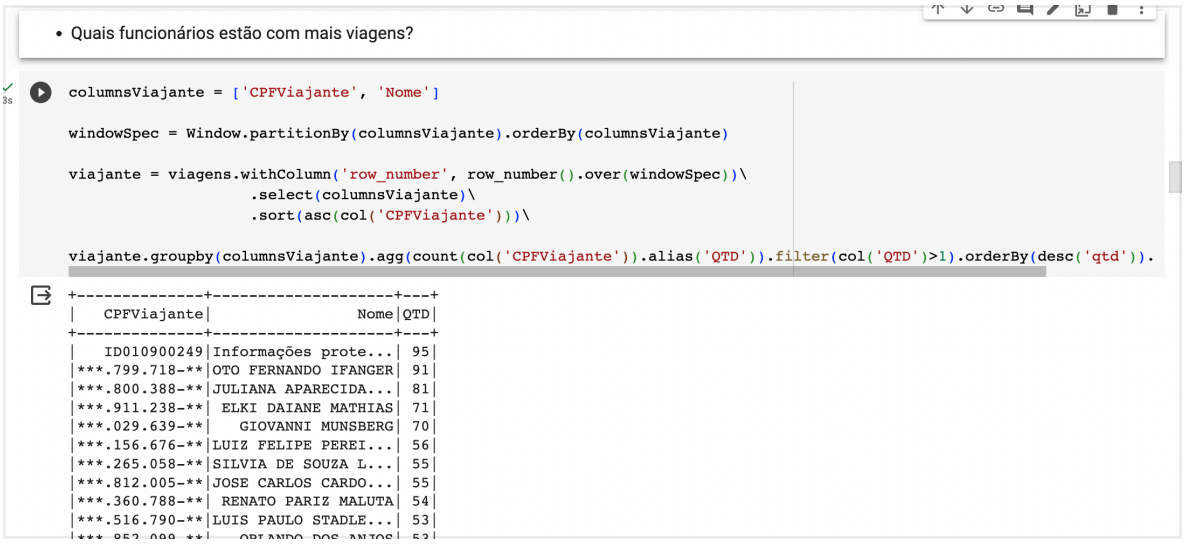
A partir de dados sobre as viagens pelos funcionários do serviço público, decidiu-se analisar:

- Qual órgão está com mais viagens?



```
+-----------------+--------------------+-----+
|CodOrgSolicitante|  NomeOrgSolicitante|  QTD|
+-----------------+--------------------+-----+
|               -1|      Sem informação|47257|
|            30108|     Polícia Federal|26092|
|            52111|Comando da Aeroná...|21460|
|            52121| Comando do Exército|16873|
|            25205|Fundação Institut...|14534|
|            30802|Polícia Rodoviári...|14494|
|            22000|Ministério da Agr...|11670|
|            20701|Instituto Brasile...| 7620|
|            37202|Instituto Naciona...| 7558|
|            44207|Instituto Chico M...| 7179|
|            36000|Ministério da Saú...| 5013|
|            52000|Ministério da Def...| 4864|
|            20000|Presidência da Re...| 3637|
|            39252|Departamento Naci...| 2961|
|            39000|Ministério dos Tr...| 2856|
|            35000|Ministério das Re...| 2548|
|            30202|Fundação Nacional...| 2507|
|            26290|Instituto Naciona...| 2485|
|            22201|Instituto Naciona...| 2095|
|            26000|Ministério da Edu...| 1873|
+-----------------+--------------------+-----+
only showing top 20 rows
```

- Quais funcionários estão com mais viagens?



- Quais cargos estão com mais viagens?

Essa pergunta não pode ser respondida, pois os dados obtidos de cargo não possuíam um atributo que permitisse o identificar unicamente. Podíamos tentar fazer essa descoberta utilizando campos estruturados, mas a quantidade de informações “Sigilosas” ou “Não informadas”atrapalhou muito na obtenção de um dado de qualidade.

Cargo	Funcao	DescricaoFuncao
NULL	OfSuperior	OfSuperior
AUXILIAR DE BIBLI...	-1	Não Informado
NULL	OfIntermed	OfIntermed
Informações prote...	Sigilosa	Informações prote...
NULL	OfIntermed	OfIntermed
NULL	OfIntermed	OfIntermed
NULL	OfIntermed	OfIntermed
NULL	-1	Não Informado
NULL	-1	Não Informado
NULL	-1	Não Informado
NULL	-1	Não Informado

Autoavaliação

- O trabalho abordado por esse MVC foi bastante desafiador para mim. Atuo na área de projetos ágeis há 7 anos e antes era desenvolvedora .net. Estava me sentindo fora do área técnica e busquei a pós em Ciência de Dados como um mecanismo de atualização. Meu conhecimento em Analytics era bem primário, o que me trouxe grande dificuldade para a realização do trabalho dessa sprint. No entanto, compreendo que a dificuldade permitiu um crescimento no que tange ao conhecimento. Consegui compreender requisitos básicos de dados, transformação e carga, utilizei um framework em phyton e comecei a ter a noção da qualidade de dados e da importância da governança deles.
- Acredito que a entrega que faço hoje possa parecer bastante básica, mas para mim representa uma evolução de aprendizado.
- Pensando em cenários futuros, pretendo aprimorar esse conhecimento usando outras ferramentas e evoluindo para um modelo não flat. Além disso, acredito ser importante refinar as questões a serem respondidas com intuito de evoluir a inteligência obtida por meio desses dados, favorecendo a tomada de decisão.

Anexo

https://colab.research.google.com/drive/1Dm_EXuLOaAkeVsB4sZdnhh-axR4KbJ7I?usp=sharing