

MVP de Engenharia de Dados

Aluno: Bruno Soares Rodrigues da Silva

Construção de um pipeline de dados utilizando tecnologias na nuvem

1. Definição do problema

O dataset objeto deste estudo refere-se a uma base de spend de uma empresa do segmento mineração, com resultados de 2018 a 2020. Entende-se como spend a relação de invoices (notas fiscais) cadastradas no ERP. A base lista as invoices de diversos países, entretanto para este exercício será premissado a análise do resultado do Brasil (que possui o spend mais representativo). O objetivo é analisar o spend das categorias de compras e a expectativa é de uma queda no resultado de 2020 devido ao impacto da pandemia de Covid-19. Estima-se que este impacto será maior nas categorias de serviço.

Contextualizando... a Covid-19 foi caracterizada como pandemia pela Organização Mundial da Saúde (OMS) em março de 2020. Muitas empresas impuseram o regime de trabalho remoto visando minimizar a propagação do vírus e, com isso, alguns contratos de fornecedores, especialmente de prestadores de serviços, tiveram que ser renegociados e/ou distratados.

Os atributos que compõem a base de dados e auxiliarão na análise estão descritos a seguir:

ATRIBUTOS	DESCRIÇÃO
YEAR	Ano de registro do documento no ERP
MONTH	Mês de registro do documento no ERP
VENDOR_CODE	ID do Fornecedor no ERP
VENDOR_NAME	Nome do Fornecedor
CONTRACT	Número do contrato no ERP
PURCHASE_ORDER	Número do Pedido de compra no ERP
CATEGORY_CODE	Código do grupo de mercadoria no ERP, que segue a metodologia UNSPSC
CATEGORY_NAME	Descrição do grupo de mercadoria
COMPANY_CODE	Código da empresa demandante
COST_TYPE	Código do tipo de custo, que definirá se trata-se de OPEX ou CAPEX
COUNTRY	País do Pedido de Compra
CURRENCY	Moeda do Pedido de compra
INCOTERM	Código da condição de frete
INVOICE_NUMBER	Número da Invoice registrada no ERP
PAYMENT_CONDITION	Código da condição de pagamento do Pedido de Compra
PLANT	Planta/Armazém do Pedido de Compra
PLANT_REGION	Estado da Planta/Armazém
PURCHASING_GROUP	Grupo de compra do gestor de categoria / comprador
PURCHASING_ORGANIZATION	Organização de compras onde foi gerado o pedido de compras
QUANTITY	Quantidade total do Pedido de Compra
SPEND	Spend Total em USD
SUPPLIER_COUNTRY	País do Fornecedor
UNIT_OF_MEASURE	Unidade de Medida do Pedido de compras
VENDOR_REGION	Estado do Fornecedor

PO_DOCUMENT_TYPE	Código no ERP que define se é uma compra de contrato ou spot
COMMODITY_GROUP	Definição da categoria de compra em Material ou Serviço

2. Carga e Análise dos Dados

Utilizou-se os recursos da Azure para upload das bases e criação do banco de dados, conforme as etapas demonstradas a seguir:

2.1. Criação da conta de armazenamento

Foram criados dois contêineres: Na pasta *insumos* temos as bases com a descrição das categorias de compras e Nome dos fornecedores

Nome	Modificado	Camada de acesso	Status do arquivo	Tipo de blob	Tamanho	Estado de concessão
db_categories.csv	13/09/2023, 14:28:39	Principal (Inferidos)		Blob de blocos	83.97 KiB	Disponível
db_vendors.csv	13/09/2023, 14:28:40	Principal (Inferidos)		Blob de blocos	606.63 KiB	Disponível

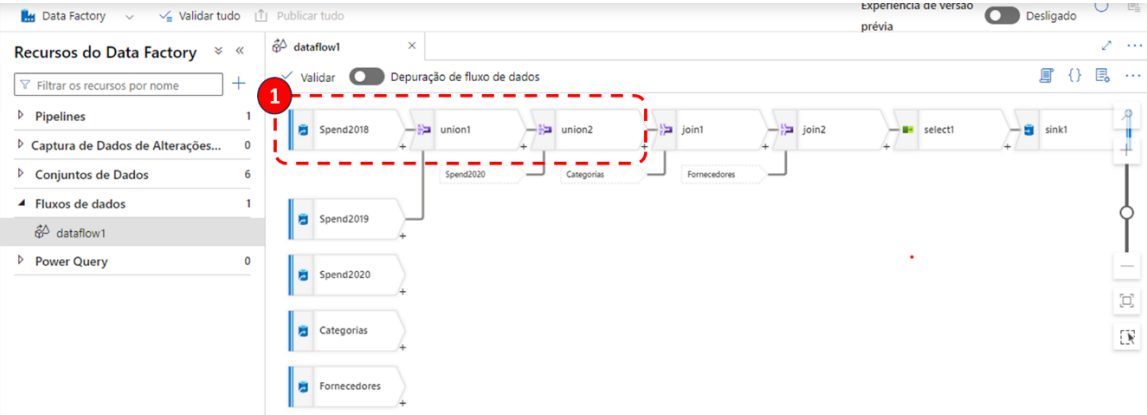
Na pasta *raw* temos as bases spend separadas por ano.

Nome	Modificado	Camada de acesso	Status do arquivo	Tipo de blob	Tamanho	Estado de concessão
spend_2018.csv	13/09/2023, 14:28:07	Principal (Inferidos)		Blob de blocos	272.63 MiB	Disponível
spend_2019.csv	13/09/2023, 14:28:10	Principal (Inferidos)		Blob de blocos	290.36 MiB	Disponível
spend_2020.csv	13/09/2023, 14:28:07	Principal (Inferidos)		Blob de blocos	293.24 MiB	Disponível

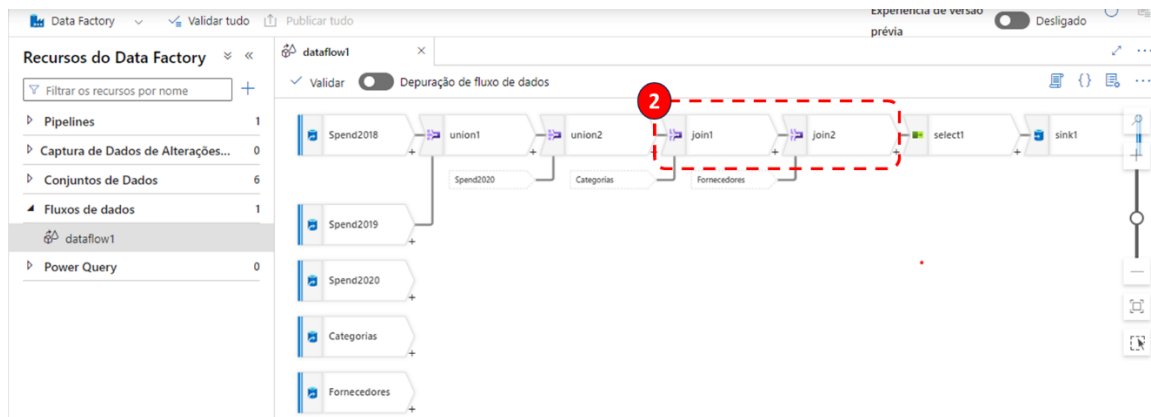
2.2. Criação do fluxo de dados e pipeline no Azure Data Factory

Utilizou-se o *dataflow* para alguns tratamentos, conforme descrito nas etapas a seguir:

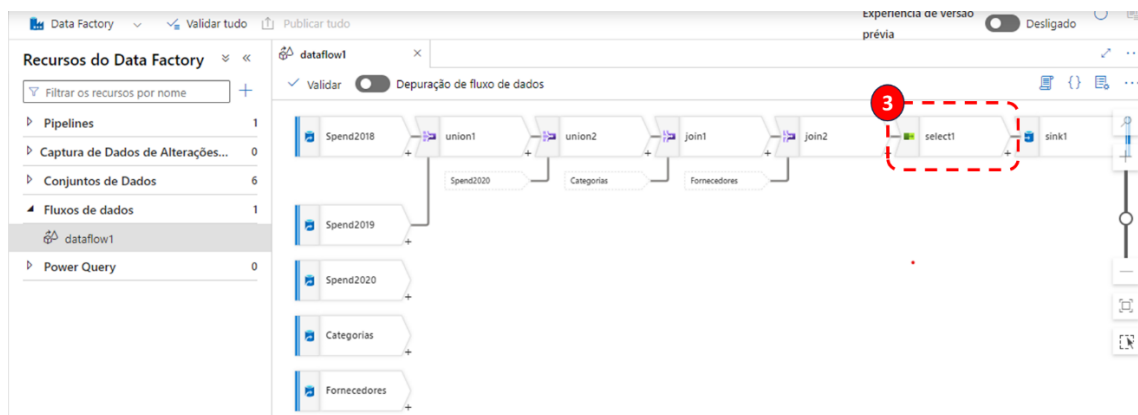
1. União dos datasets 2018, 2019 e 2020



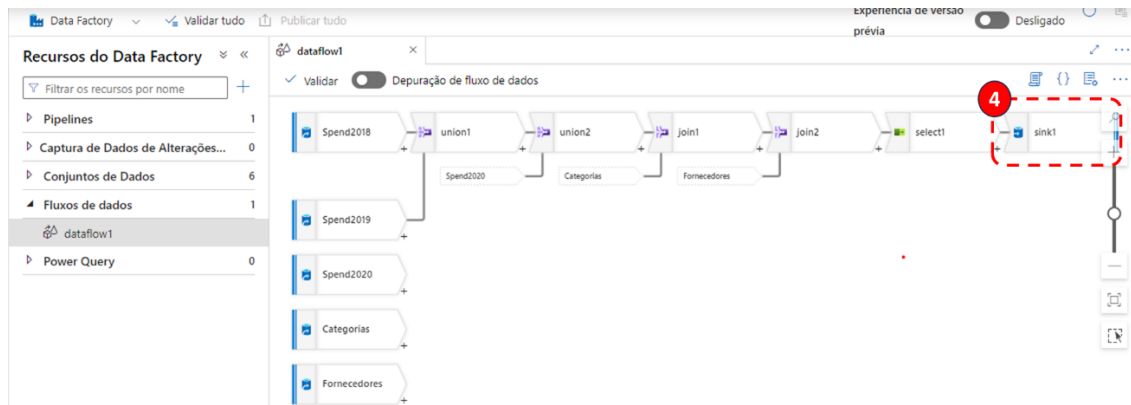
2. *Join* com as bases Categorias, para trazer a descrição e Fornecedores para trazer o nome dos mesmos



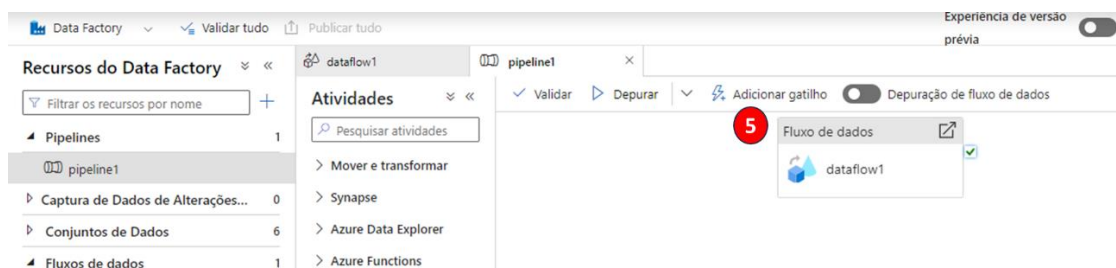
3. Exclusão de algumas colunas



4. Subir os dados para o SQL Database



5. Validação e depuração do fluxo de dados



2.3. Executando consultas no *database* através do Azure Data Studio.

Pesquisando as primeiras 20 linhas do *database*

```
[2] 1 SELECT TOP 20 * FROM dbo.spend
```

(20 rows affected)

Total execution time: 00:00:00.166



	YEAR	MONTH	VENDOR_CODE	VENDOR_NAME	CONTRACT	PURCHASE_ORDER	CATEGORY_CODE
1	2018	1	90000237	COFERMETA SA	NULL	4100280049	40161534A
2	2018	1	90000237	COFERMETA SA	NULL	4100282565	31162400
3	2018	1	90000237	COFERMETA SA	NULL	4100282858	40142300
4	2018	1	90000237	COFERMETA SA	NULL	4100284329	40141660A
5	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100138003	39121732A
6	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100138003	EP011601
7	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100276976	31161627
8	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100276976	31161800
9	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100276976	39122325
10	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100277619	11111111C
11	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100277619	26120000
12	2018	1	90000240	IMPORTADORA ELETRICA BRASILIA LTDA	NULL	4100277619	39121106

Iniciaremos a análise do spend do Brasil pelo número de Pedidos de Compra das categorias de material e serviço emitidas por ano:

Pesquisar o total de Pedidos de compra de Material e Serviço emitidos para o Brasil por ano.

```
1 SELECT YEAR, COUNT(DISTINCT PURCHASE_ORDER) AS Total_PO, COMMODITY_GROUP
2 FROM dbo.spend
3 WHERE COUNTRY = 'Brazil'
4 GROUP BY YEAR, COMMODITY_GROUP
5 ORDER BY YEAR
```



SQL

(6 rows affected)

Total execution time: 00:01:46.054



	YEAR	Total_PO	COMMODITY_GROUP
1	2018	520516	MATERIAL
2	2018	279811	SERVICO
3	2019	567045	MATERIAL
4	2019	333600	SERVICO
5	2020	557491	MATERIAL
6	2020	353789	SERVICO

Pode-se verificar um aumento do número de pedidos de compra de serviço e diminuição do número de pedidos de compra de material. Vamos fazer a mesma análise com valor do spend para verificar se isso se mantém.

Avaliar o resultado do spend das categorias de material e serviço no Brasil por ano

```
[12] 1 SELECT YEAR, SUM(SPEND) AS Total_Spend, COMMODITY_GROUP
2 FROM dbo.spend
3 WHERE COUNTRY = 'Brazil'
4 GROUP BY YEAR, COMMODITY_GROUP
5 ORDER BY YEAR
```

SQL

(6 rows affected)

Total execution time: 00:00:17.144



	YEAR	Total_Spend	COMMODITY_GROUP
1	2018	4389713939	MATERIAL
2	2018	3350576366	SERVICO
3	2019	4501405970	MATERIAL
4	2019	4156972333	SERVICO
5	2020	4032009602	MATERIAL
6	2020	4516621738	SERVICO

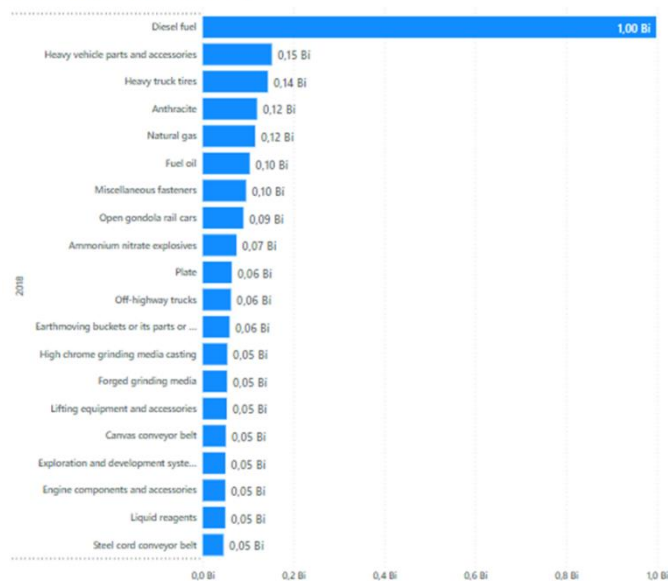
Nesta consulta foi possível identificar que, ao contrário da expectativa inicial de impacto no spend das categorias de serviço, tivemos uma queda no spend das categorias de material em 2020 (de USD 4,5 Bi em 2019 para USD 4,0 Bi em 2020). Isso nos leva a entender que os contratos de serviço que foram distratados não tinham grande representativa no spend da categoria. Outro ponto seria uma renegociação da condição de homem/hora trabalhada.

E um dos cenários possíveis que justifiquem a queda do spend de material seria uma renegociação de prazo de entrega de material, considerando que as operações estavam operando em regime de escala para diminuição do efetivo e, consequentemente, minimizar os riscos de contaminação... ou até mesmo cancelamento de pedidos que não seriam essenciais.

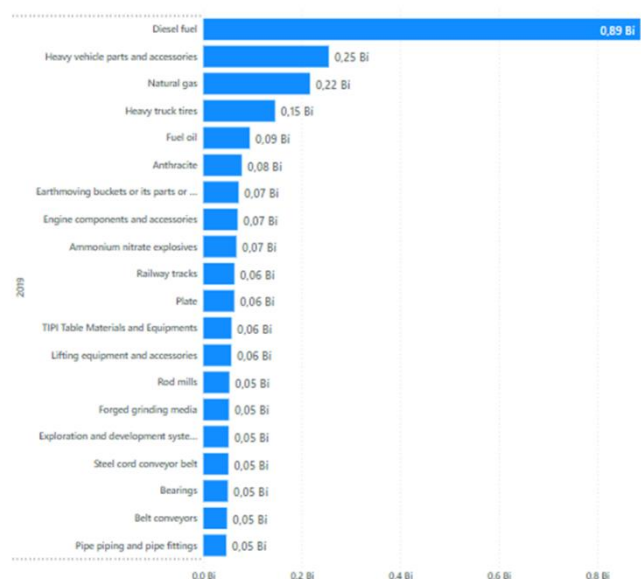
2.4. Análise das categorias de compras através do PowerBI

Usamos o powerbi como recurso para visualizarmos o comportamento das top 20 categorias de material no período 2018 – 2020 e, desta forma, colhermos algum insight.

SPEND USD 2018 - MATERIAL - TOP 20



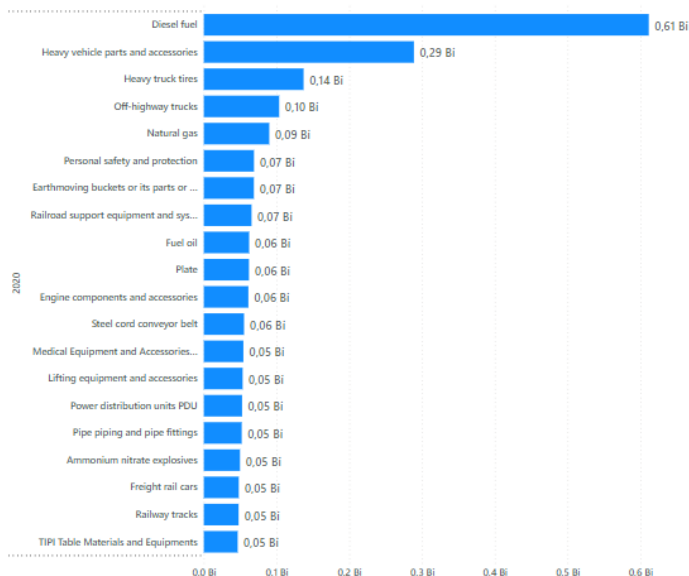
SPEND USD 2019 - MATERIAL - TOP 20



O spend das categorias de material entre 2018 e 2019 teve variação entre as top 20, mas tudo dentro do esperado, com as categorias de combustível mantendo-se no topo.

Vamos analisar o resultado de 2020.

SPEND USD 2020 - MATERIAL - TOP 20



Podemos observar uma grande queda do spend de Diesel, indo de USD 8,9Mi em 2019 para USD 6,1Mi em 2020... entretanto o que chamou a atenção foi o spend da categoria *Personal safety and protection* que teve um spend de USD 7Mi, sendo a 6ª categoria de maior spend neste ano, considerando que nos anos anteriores a mesma não configurou no top 20. Isso faz total sentido, considerando o cenário de pandemia em 2020.

3. Considerações Finais

Este trabalho buscou analisar o spend das categorias de compras do Brasil no período de 2018 a 2020 e verificar um possível impacto da pandemia de Covid-19 no resultado de 2020. A expectativa era de uma diminuição do spend das categorias de serviço visto que, devido ao cenário de imposição do trabalho remoto nos escritórios para minimizar o risco de propagação do vírus, assim como o estabelecimento de turnos nas operações com a mesma finalidade... muitos contratos foram renegociados ou distratados. Entretanto o que se observou foi uma queda de USD 500Mi no spend das categorias de materiais. O diesel, que é a categoria mais representativa, teve uma queda de USD 2,8Mi de 2019 para 2020... e a categoria de Segurança e Proteção pessoal que, não figurava entre as 20 categorias de maior spend em 2018 e 2019, teve grande representatividade em 2020, com um spend de USD 7Mi.

Como próximos passos para este trabalho, fica a sugestão de análise das categorias nos anos pós pandemia para avaliação do comportamento... assim como estender este mesmo exercício para outros países, a fim de comparação.