

PUC - MVP da disciplina Engenharia de Dados

Objetivo

Utilizando os dados de imigrações para o Canada. Analisar:

1. Quais os países que mais migram para o Canada?
2. O número de imigrantes para o Canada tem aumentado?
3. A imigração vem majoritariamente de países desenvolvidos?

Detalhamento:

O Dataset "Immigration to Canada" foi extraído do Kaggle para a máquina local.



<https://www.kaggle.com/datasets/ammaraahmad/immigration-to-canada>

Modelagem

O dado será modelado como um Data Lake, fazendo a carga da planilha de 39 Colunas e 193 linhas.

Catálogo de dados:

Colunas	Descrição	Type	Min	Max
Country	Páís de Origem dos Imigrantes	String	-	-
Continent	Continente de Origem dos Imigrantes	String	-	-
Region	Região de Origem dos Imigrantes	String	-	-
DevName	Se o país é desenvolvido ou se está em desenvolvimento	String	-	-
1980	Numero de Imigrantes no ano de 1980	int	0	50000
1981	Numero de Imigrantes no ano de 1981	int	0	50000
1982	Numero de Imigrantes no ano de 1982	int	0	50000
...
...
2013	Numero de Imigrantes no ano de 2013	int	0	50000
Total	Numero de Imigrantes somando todos os anos	int	0	100000

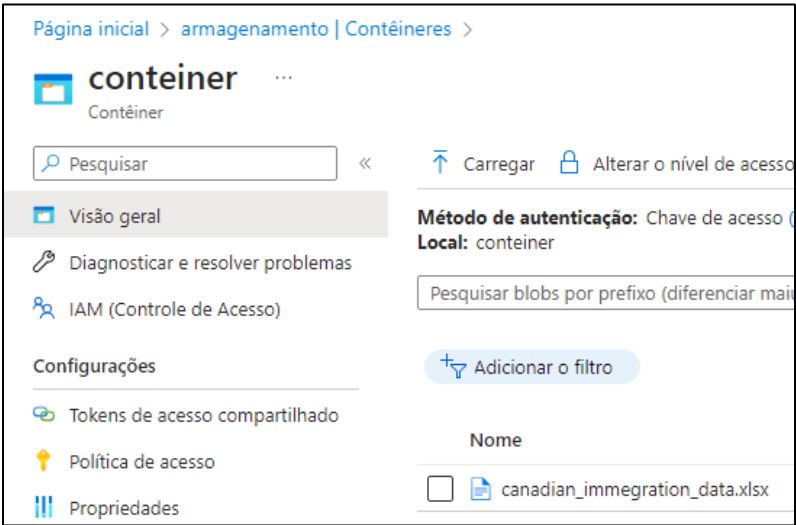
Carga

Para ser realizado a carga dos dados foi escolhido a plataforma em nuvem “Microsoft Azure”, onde criei uma conta.

Dentro da conta criei um grupo de recursos para poder colocar todos os próximos recursos que irei precisar criar:



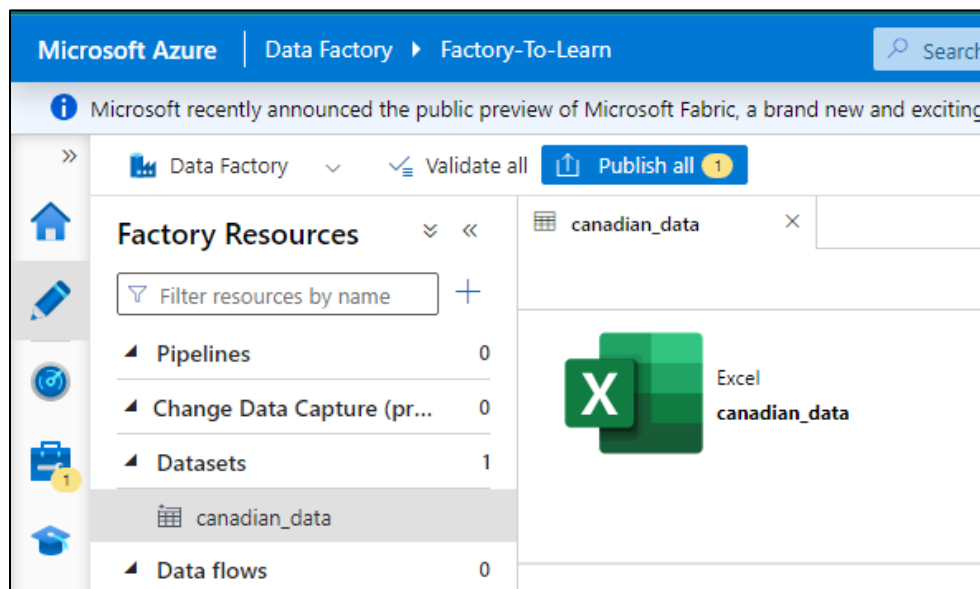
Após a criação do recurso criei uma “Conta de armazenamento” dentro criei um contêiner onde fiz upload da minha planilha.



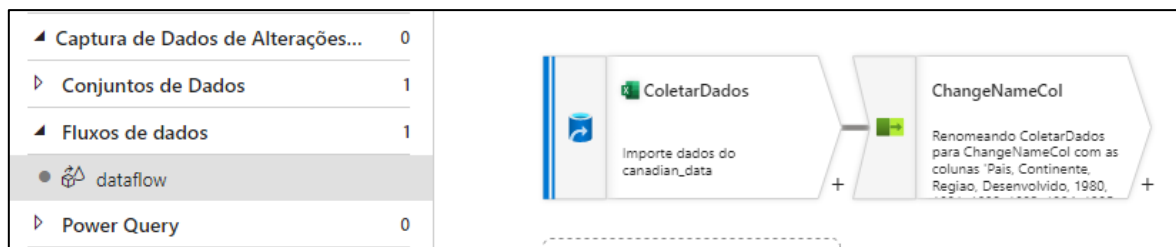
Criei também um servidor para poder colocar o banco de dados e dentro do servidor criei um banco de dados e depois criei mais um recurso, o data Factory, após criar todos os recursos ficou assim minha lista de recursos:

<input type="checkbox"/> Nome ↑↓	Tipo ↑↓
<input type="checkbox"/> armazenamento	Conta de armazenamento
<input type="checkbox"/> bacodedados (servidordadossql/bacodedados)	Banco de dados SQL
<input type="checkbox"/> Factory-To-Learn	Data factory (V2)
<input type="checkbox"/> servidordadossql	SQL Server

Dentro do Data Factory adicionei um Dataset usando o “Azure Blob Storage” configurei para ele coletar os dados da planilha dentro do contêiner criado anteriormente.



Depois criei um Data Flow onde o primeiro bloco faz a coleta do dataset já definindo o formato do dado de cada coluna e o segundo troca o nome das colunas.



<input type="checkbox"/>	Coluna do ColetarDados		Nomear como	
<input type="checkbox"/>	abc Country	→	Pais	+
<input checked="" type="checkbox"/>	abc Continent	→	Continente	+
<input type="checkbox"/>	abc Region	→	Regiao	+
<input type="checkbox"/>	abc DevName	→	Desenvolvido	+
<input type="checkbox"/>	123 1980	→	1980	+

Depois criei o terceiro bloco para salvar o dataset tratado no banco de dados.

ColetarDados

Importe dados do canadian_data

ChangeNameCol

Renomeando ColetarDados para ChangeNameCol com as colunas: Pais, Continente, Regiao, Desenvolvido, 1980,

Salvar

Colunas: 39 total

Coletor

Configurações

Erros

Mapeamento

Otimizar

Inspeccionar

Visualização de dados

Tipo de coletor *

Conjunto de Dados

Embutido

Cache

Conjunto de Dados *

AzureSqlTable1

Testar conexão

Abrir

Novo

Opções

☒ Permitir o descompartilhamento de esquema ⓘ

☒ Validar esquema ⓘ

Depois foi criado um pipeline onde foi inserido o fluxo de dados. O pipeline foi executado e obteve sucesso.

Fluxo de dados

dataflow

Parâmetros

Variáveis

Configurações

Saída

ID de execução de pipeline: 468e729c-5e67-4519-8bd2-852b7ec8deb2 ⓘ

Status do pipeline

All status ▾

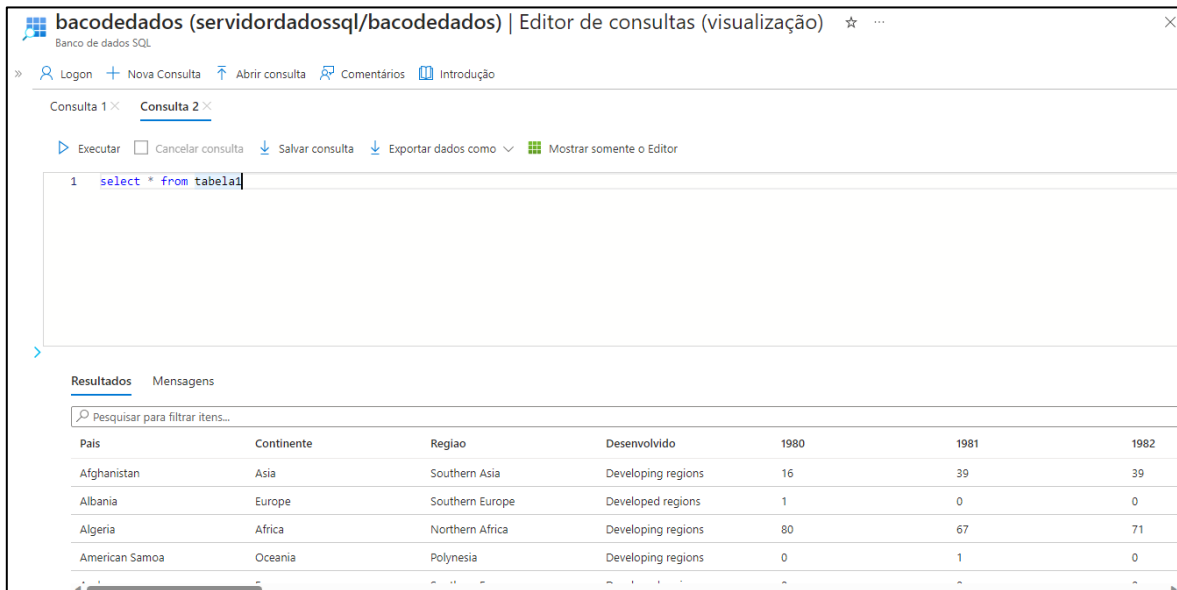
Monitorar nas Métricas do pipeline

Mostrando 1 - 1 de 1 itens

Activity name	Activity status	Run start	Duration
dataflow	✓ Bem-sucedido	9/13/2023, 7:17:58 PM	52s

Análise (Qualidade dos dados)

Para Analisar os dados foi solicitado o `select * from tabela1`, para visualizar os dados como o Dataset não possui muitas linhas uma visualização completa da tabela já foi o suficiente para analisar a qualidade dos dados.



The screenshot shows a web-based SQL editor interface. At the top, the title bar reads "bacodedados (servidordadossql/bacodedados) | Editor de consultas (visualização)". Below the title bar, there are navigation links: "Logon", "Nova Consulta", "Abrir consulta", "Comentários", and "Introdução". The main area is divided into two tabs: "Consulta 1" and "Consulta 2". The "Consulta 2" tab is active, showing a SQL query: `1 select * from tabela1`. Below the query editor, there are buttons: "Executar", "Cancelar consulta", "Salvar consulta", "Exportar dados como", and "Mostrar somente o Editor". The "Executar" button is highlighted. Below the buttons, there is a section for "Resultados" and "Mensagens". The "Resultados" section is active, showing a table with 7 columns: "Pais", "Continente", "Regiao", "Desenvolvido", "1980", "1981", and "1982". The table contains 5 rows of data.

Pais	Continente	Regiao	Desenvolvido	1980	1981	1982
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39
Albania	Europe	Southern Europe	Developed regions	1	0	0
Algeria	Africa	Northern Africa	Developing regions	80	67	71
American Samoa	Oceania	Polynesia	Developing regions	0	1	0

Análise (Solução do Problema)

Nesta etapa irei realizar as consultas em SQL para obter as respostas das perguntas do início do projeto.

1. Quais os países que mais migram para o canada?

Utilizando Esta Query: `select pais, total from tabela1 order by total desc`

Obtive a Resposta Abaixo:

pais	total
India	691904
China	659962
United Kingdom	551500
Philippines	511391
Pakistan	241600
United States of America	241122

2. O número de imigrantes para o canada tem aumentado?

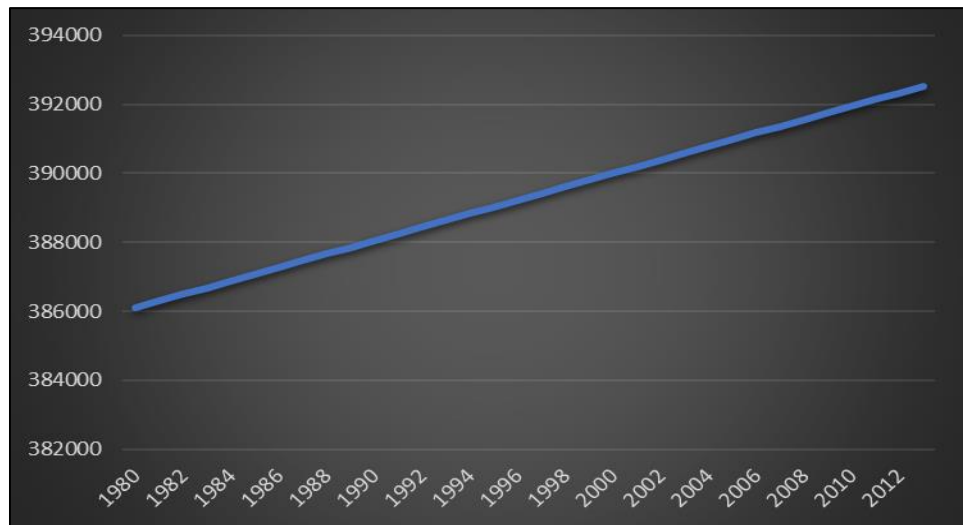
Utilizando Esta Query:

```
“select  
    sum(1980) as '1980', sum(1981) as '1981', sum(1982) as '1982',  
    sum(1983) as '1983', sum(1984) as '1984', sum(1985) as '1985',  
    sum(1986) as '1986', sum(1987) as '1987', sum(1988) as '1988',  
    sum(1989) as '1989', sum(1990) as '1990', sum(1991) as '1991',  
    sum(1992) as '1992', sum(1993) as '1993', sum(1994) as '1994',  
    sum(1995) as '1995', sum(1996) as '1996', sum(1997) as '1997',  
    sum(1998) as '1998', sum(1999) as '1999', sum(2000) as '2000',  
    sum(2001) as '2001', sum(2002) as '2002', sum(2003) as '2003',  
    sum(2004) as '2004', sum(2005) as '2005', sum(2006) as '2006',  
    sum(2007) as '2007', sum(2008) as '2008', sum(2009) as '2009',  
    sum(2010) as '2010', sum(2011) as '2011', sum(2012) as '2012',  
    sum(2013) as '2013'  
from tabela1  
”
```

E obtive esta resposta:

1980	1981	1982	1983	1984	1985	1986
386100	386295	386490	386685	386880	387075	387270

Como esta ferramenta que estou utilizando não tem acesso fácil ao python, copiei o resultado e coloquei no Excel para poder gerar um gráfico e visualizar esses dados, obtive a seguinte resposta:



Logo sim os imigrantes do canada têm aumentado de forma bem linear ao longo dos anos

3. A imigração vem majoritariamente de países desenvolvidos?

Utilizando esta Query:

```
“select
    Desenvolvido,
    sum(total) as 'Valor Total'
from tabela1
    group by Desenvolvido
    order by 'valor total' desc
”
```

E obtive esta resposta:

Desenvolvido	Valor Total
Developing regions	4695142
Developed regions	1714011

Logo a Imigração vem majoritariamente de países em desenvolvimento.