

**MVP SPRINT 1 – ENGENHARIA DE DADOS**  
**ANÁLISE DE DADOS DOS BILIONÁRIOS NO MUNDO**

JORGE DE FIGUEIREDO COSTA JUNIOR

RIO DEJANEIRO  
2023

## 1-Introdução

Este projeto tem como objetivo principal explorar a Engenharia de Dados aplicada à análise de uma base de dados intrigante: a lista de bilionários do mundo. O projeto propõe a criação de armazenamento em nuvem e o desenvolvimento de consultas SQL para encontrarmos respostas sobre esses dados.

Vamos passar pela criação do armazenamento, tratamento dos dados e consultas para buscarmos respostas para algumas perguntas interessantes.

## 2- Objetivos

A lista de bilionários do mundo é um recurso rico em informações que contém dados sobre as pessoas mais influentes no cenário econômico global.

Neste projeto vamos em busca das seguintes respostas:

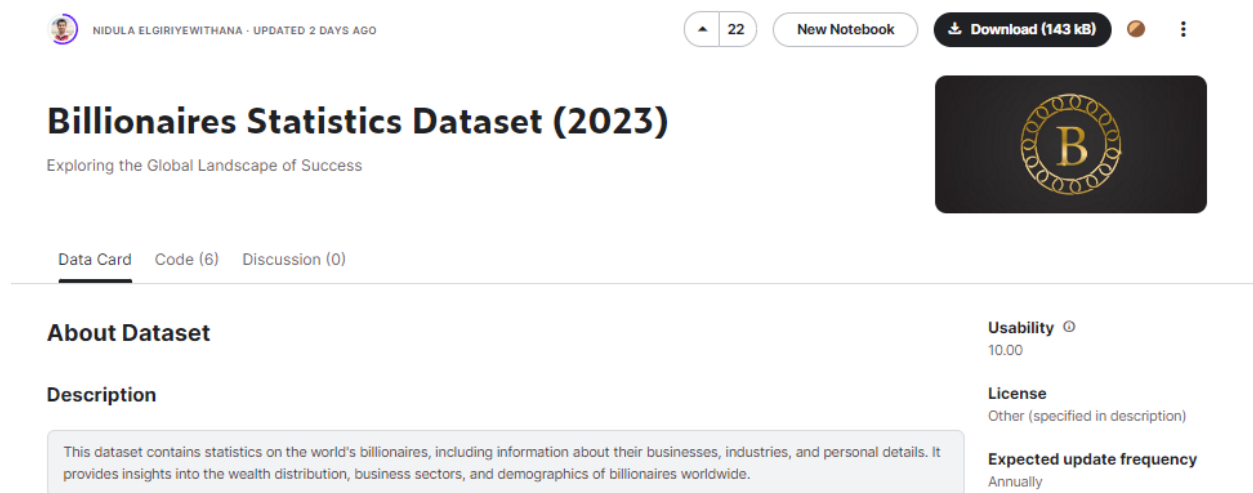
- 1) Quantidade total de bilionarios
- 2) Quantidade total por gênero
- 3) Quantidade total por categoria
- 4) Quantidade de brasileiros na lista
- 5) Quantidade de brasileiros por gênero
- 6) Quantidade de bilionários brasileiros por tecnologia
- 7) Bilionários brasileiros por idade

## 3- Especificação Técnica

### 3.1. Coleta de Dados

A base analisada foi retirada do seguinte caminho:

<https://www.kaggle.com/datasets/nelgiriyeewithana/billionaires-statistics-dataset>



**Billionaires Statistics Dataset (2023)**  
Exploring the Global Landscape of Success

**About Dataset**

**Description**

This dataset contains statistics on the world's billionaires, including information about their businesses, industries, and personal details. It provides insights into the wealth distribution, business sectors, and demographics of billionaires worldwide.

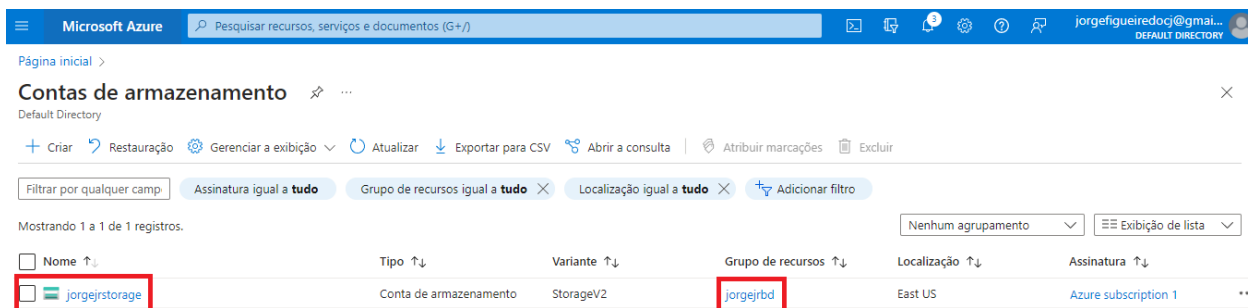
**Usability** 10.00

**License** Other (specified in description)

**Expected update frequency** Annually

Os dados foram baixados para uma máquina local e inseridos manualmente em uma conta de armazenamento do Microsoft Azure:

- Criado storage



**Microsoft Azure** | Pesquisar recursos, serviços e documentos (G+/I)

Página inicial >

**Contas de armazenamento** | Default Directory

+ Criar | Restauração | Gerenciar a exibição | Atualizar | Exportar para CSV | Abrir a consulta | Atribuir marcações | Excluir

Filtrar por qualquer campo: Assinatura igual a tudo | Grupo de recursos igual a tudo | Localização igual a tudo | Adicionar filtro

Mostrando 1 a 1 de 1 registros.

Nome ↑↓	Tipo ↑↓	Variante ↑↓	Grupo de recursos ↑↓	Localização ↑↓	Assinatura ↑↓
jorgejrstorage	Conta de armazenamento	StorageV2	jorgejrbd	East US	Azure subscription 1

- Criado container

**jorgejrstorage** | Contêineres ✨ ☆ ...

Conta de armazenamento

Pesquisar

Eventos

Navegador de armazenamento

Migrador de Armazenamento do Azure

Armazenamento de dados

Contêineres

+ Contêiner Alterar o nível de acesso Restaurar contêineres Atualizar Excluir ...

Pesquisar contêineres por prefixo

Mostrar contêineres excluídos

Nome	Última modificação	Nível de acesso anô...	Estado de concessão
<input type="checkbox"/> \$logs	01/10/2023, 13:54:03	Privado	Disponível ***
<input type="checkbox"/> jorgejrcontainer	01/10/2023, 14:02:46	Privado	Disponível ***

- Arquivo inserido manualmente

**jorgejrcontainer** ...

Contêiner

Pesquisar

Carregar Alterar o nível de acesso Atualizar Excluir Alterar a camada Adquirir concessão Interromper concessão ...

Método de autenticação: Chave de acesso (Alternar para a Conta de Usuário do Azure AD)

Local: jorgejrcontainer

Pesquisar blobs por prefixo (diferenciar maiúsculas de minúsculas)

Mostrar blobs excluídos

Adicionar o filtro

Nome	Modificado	Camada de acesso	Status do arquivo	Tipo de blob	Tamanho	Estado de concessão
<input type="checkbox"/> Billionaires_Statistics...	01/10/2023, 14:03:02	Principal (Inferidos)		Blob de blocos	675.57 KiB	Disponível ***

## 3.2. Modelagem dos dados

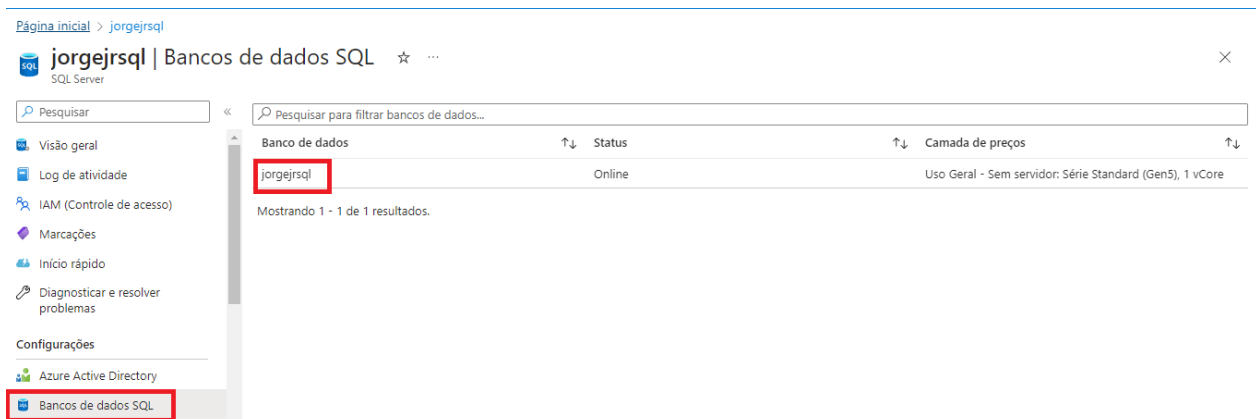
Ordenar	Coluna	Tipo
1	RANK	string
2	FINAL_WORTH	string
3	CATEGORY	string
4	PERSON_NAME	string
5	AGE	string
6	COUNTRY	string
7	CITY	string
8	SOURCE	string
9	INDUSTRIES	string
10	COUNTRY_OF_CITIZENSHIP	string
11	ORGANIZATION	string
12	STATUS	string
13	GENDER	string

RANK	int	A classificação do bilionário em termos de riqueza
FINAL_WORTH	int	O patrimônio líquido final do bilionário em dólares americanos
CATEGORY	string	A categoria ou setor em que opera o negócio do bilionário
PERSON_NAME	string	O nome completo do bilionário
AGE	int	A idade do bilionário
COUNTRY	string	O país em que o bilionário reside
CITY	string	A cidade em que o bilionário reside
SOURCE	string	A fonte da riqueza do bilionário
INDUSTRIES	string	As indústrias associadas aos interesses comerciais do bilionário
COUNTRY_OF_CITIZENSHIP	string	O país de cidadania do bilionário
ORGANIZATION	string	O nome da organização ou empresa associada ao bilionário
STATUS	string	"D" representa bilionários que se fizeram sozinhos (fundadores/empreendedores) e "U" indica riqueza herdada ou não conquistada
GENDER	string	O gênero do bilionário

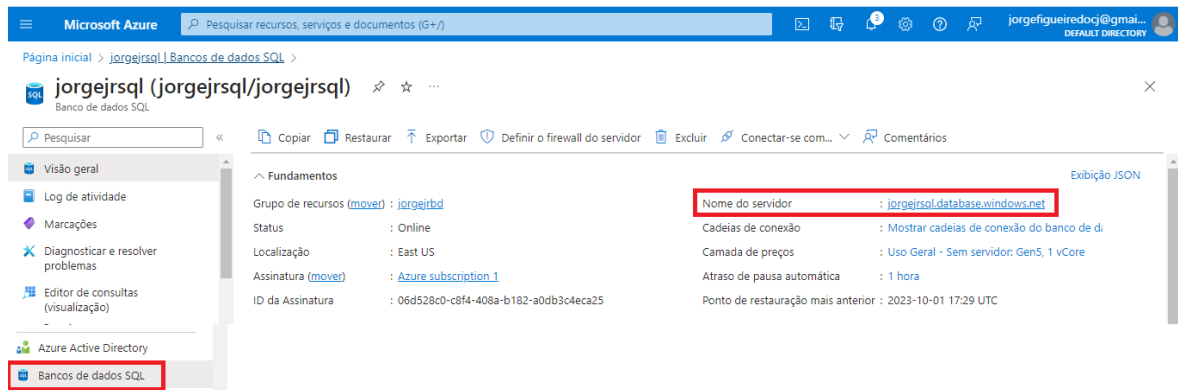
### 3.3. Carga dos Dados

O ETL foi realizado pelo Azure Data Factory

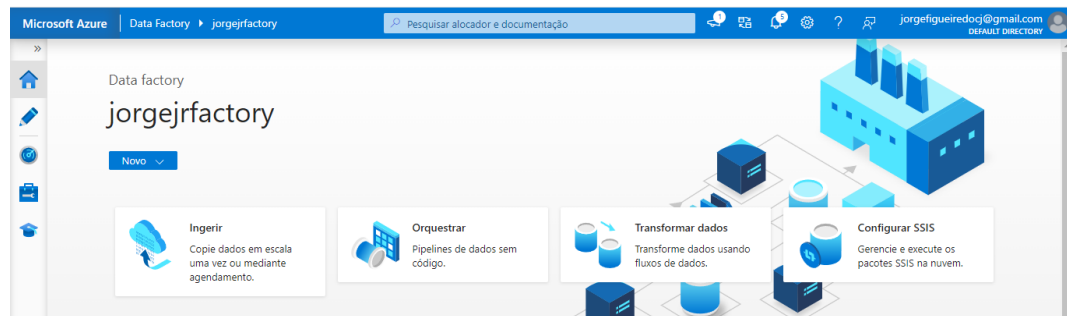
- Criado banco de dados SQL



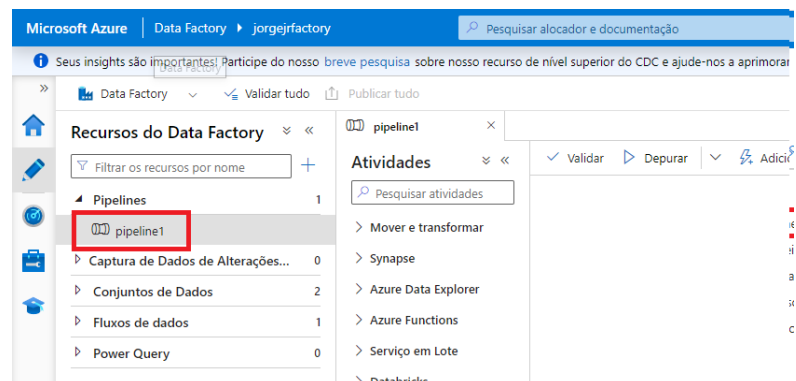
- Criado servidor SQL com autenticação por usuário e senha



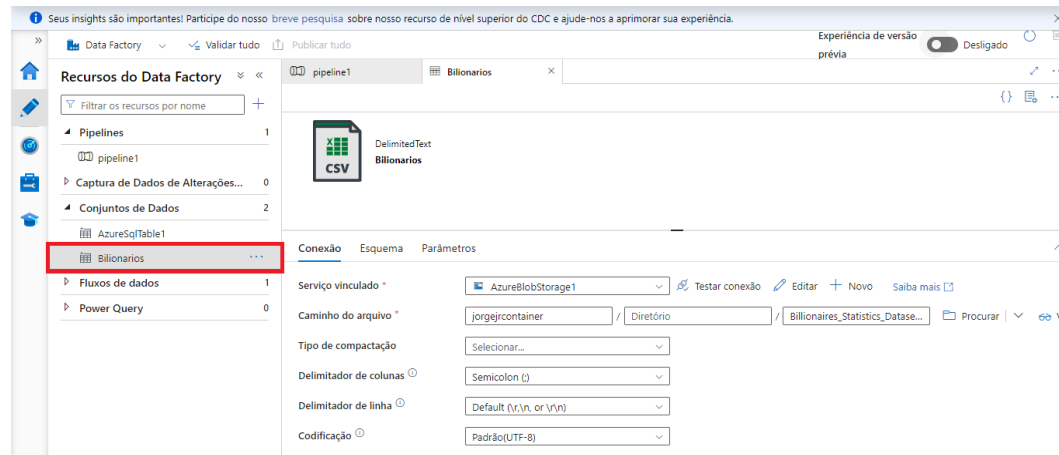
- Criada instância no Data Factory



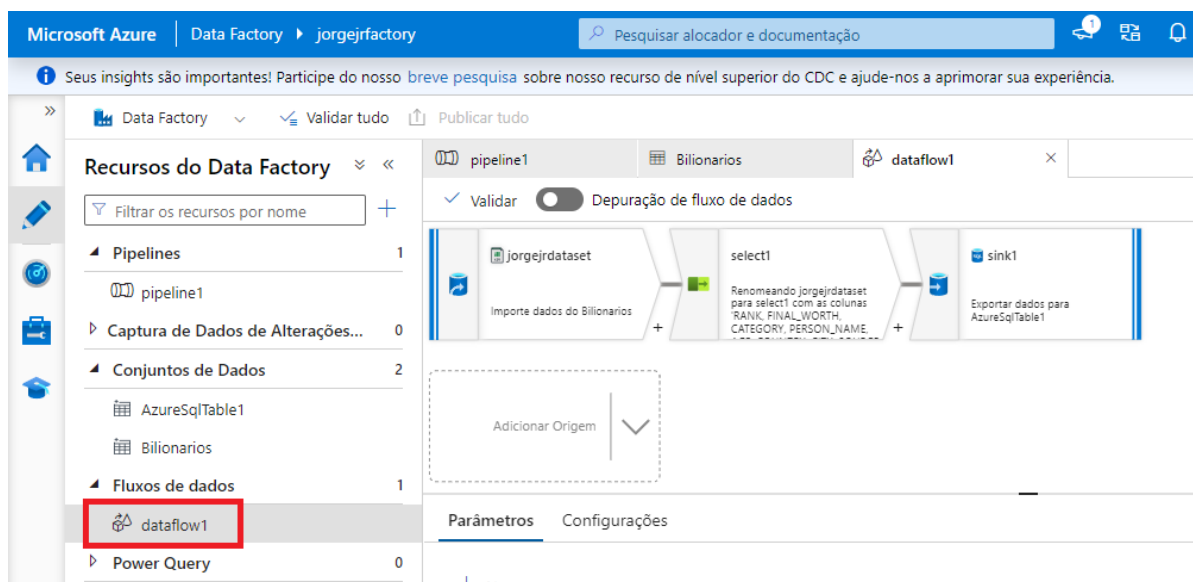
- Criado pipeline



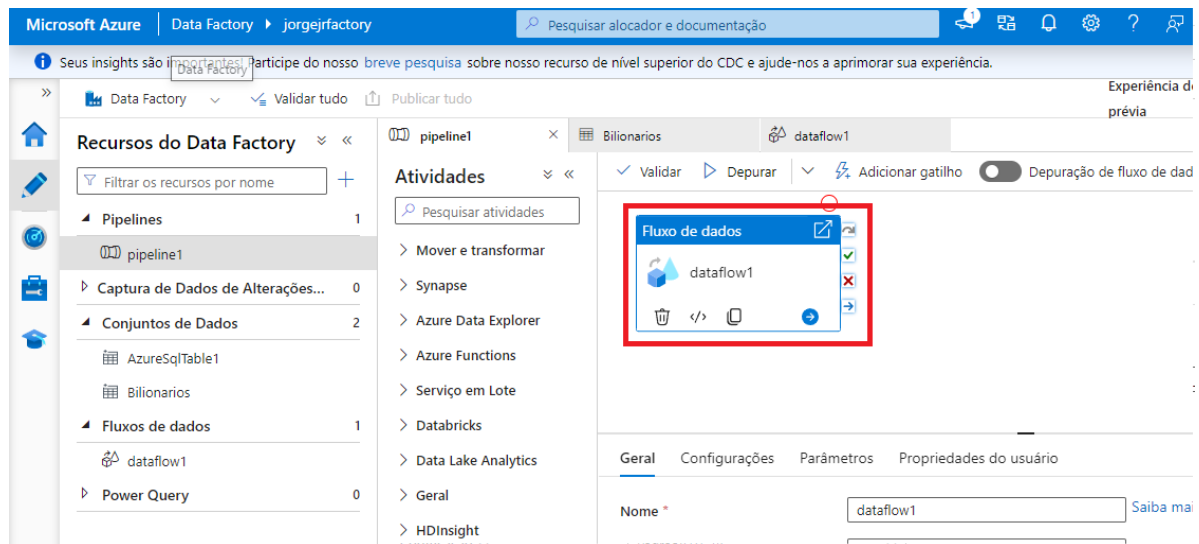
- Criado Banco de Dados para leitura do arquivo no Azure Blob Storage. Foi necessário criar um link para comunicação.



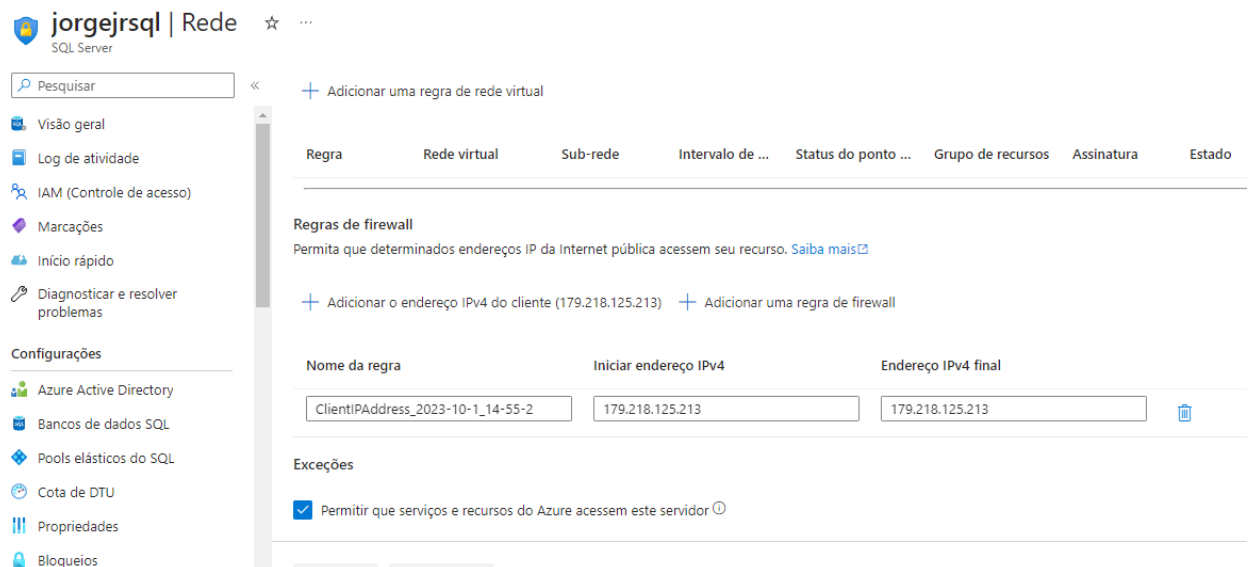
- Criando dataflow
  - Adicionado Dataset e inserida as informações dos dados de entrada
  - Adicionado Select para retirada de 22 colunas que não eram interessantes para a análise
  - Adicionado sink para conexão com o banco de dados SQL



- Foi adicionado o Dataflow dentro do Pipeline

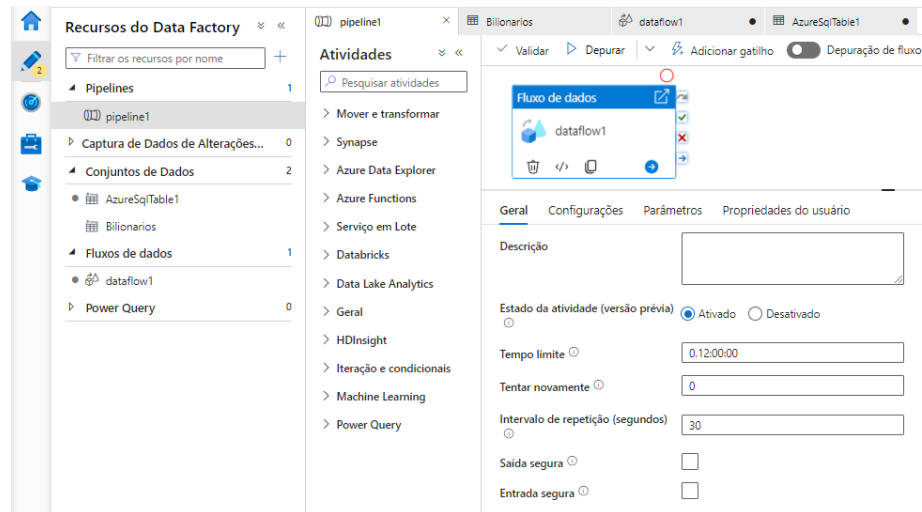


Foi necessário configurar acesso ao Banco SQL para acesso do Pipeline e acesso externo de consulta ao Banco.





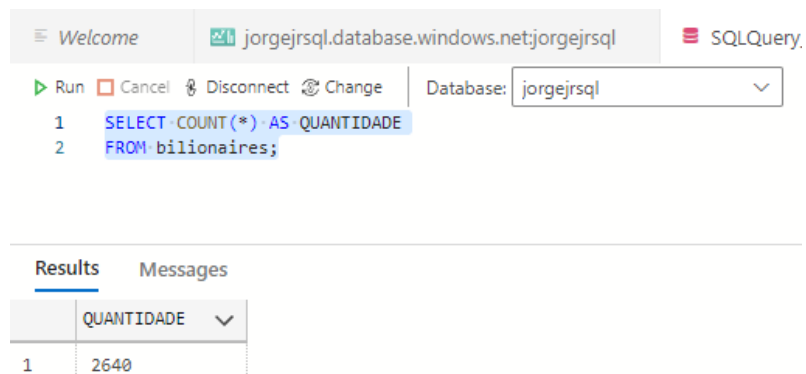
- Pipeline executado com sucesso



## 4- Solução do problema

Para consulta e análise da base de dados, foi baixado o aplicativo Azure Data Studio e conectado a base **jorgejrsql.database.windows.net**

### 4.1. Quantidade total de Bilionarios



## 4.2. Diversidade por gênero

Welcome

jorgejrsql.database.windows.net:jorgejrsql

SQLQuery\_1 - jorgej...radmin)

Run

Cancel

Disconnect

Change

Database: jorgejrsql

Estimated Plan

Enable

```

1  select GENDER, COUNT (*) AS QUANTIDADE,
2  CAST(COUNT(*) * 100.0 / SUM(COUNT(*) OVER () AS DECIMAL(10, 2)) AS PORCENTAGEM
3  from billionaires
4  GROUP BY GENDER;
```

Results

Messages

	GENDER	QUANTIDADE	PORCENTAGEM
1	F	337	12.77
2	M	2303	87.23

Através dessa consulta podemos verificar que a diferença entre os gêneros dos bilionários do mundo é muito alta e mostra a realidade que vivemos hoje, deixando claro o gap imposto pela sociedade até os dias de hoje.

## 4.3. Quantidade de Bilionários por categoria

Run

Cancel

Disconnect

Change

Database: jorgejrsql

Estimated Plan

Enable Actual

```

1 SELECT DISTINCT CATEGORY, COUNT(*) AS QUANTIDADE
2 FROM billionaires
3 GROUP BY CATEGORY
4 ORDER BY QUANTIDADE DESC;
```

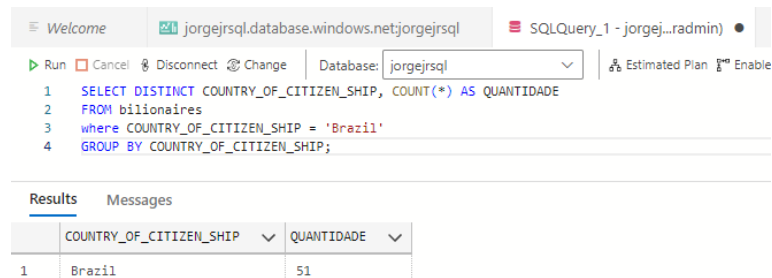
Results

Messages

	CATEGORY	QUANTIDADE
1	Finance & Investments	372
2	Manufacturing	324
3	Technology	314
4	Fashion & Retail	266
5	Food & Beverage	212
6	Healthcare	201
7	Real Estate	193
8	Diversified	187
9	Energy	100
10	Media & Entertainment	91
11	Metals & Mining	74
12	Automotive	73
13	Service	53
14	Construction & Engineering	45
15	Logistics	40
16	Sports	39
17	Telecom	31
18	Gambling & Casinos	25

Analizando os bilionários por categoria, chegamos a constatação que as instituições financeiras dominam o ranking, mas vemos o ramo de tecnologia se aproximando e precisamos avaliar como será o futuro.

#### 4.4. Quantidade Bilionários brasileiros

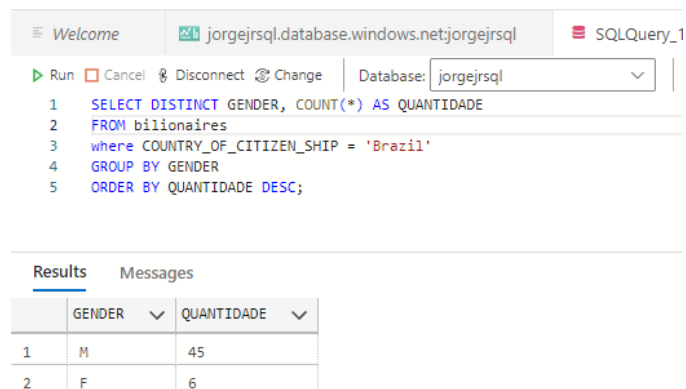


The screenshot shows the SQLQuery\_1 window in jorgejrsql. The query is: `SELECT DISTINCT COUNTRY_OF_CITIZEN_SHIP, COUNT(*) AS QUANTIDADE FROM bilionaires where COUNTRY_OF_CITIZEN_SHIP = 'Brazil' GROUP BY COUNTRY_OF_CITIZEN_SHIP;` The results table has two columns: COUNTRY\_OF\_CITIZEN\_SHIP and QUANTIDADE. The first row shows 'Brazil' with a count of 51.

	COUNTRY_OF_CITIZEN_SHIP	QUANTIDADE
1	Brazil	51

Em um país com dimensões continentais temos um total de 51 brasileiros nessa lista.

#### 4.5. Quantidade Bilionários brasileiros divididos por gênero

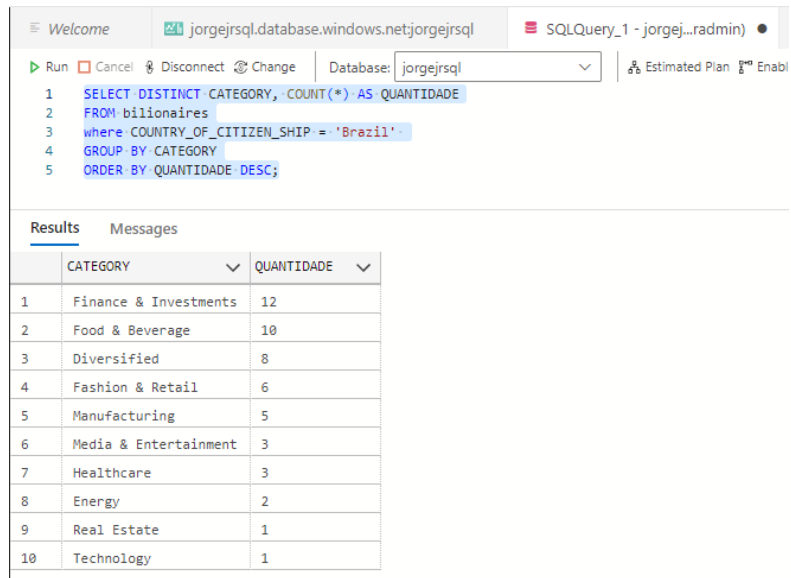


The screenshot shows the SQLQuery\_1 window in jorgejrsql. The query is: `SELECT DISTINCT GENDER, COUNT(*) AS QUANTIDADE FROM bilionaires where COUNTRY_OF_CITIZEN_SHIP = 'Brazil' GROUP BY GENDER ORDER BY QUANTIDADE DESC;` The results table has two columns: GENDER and QUANTIDADE. The first row shows 'M' with a count of 45, and the second row shows 'F' with a count of 6.

	GENDER	QUANTIDADE
1	M	45
2	F	6

Mantendo a ocorrência identificada quando a consulta foi feita para o mundo, há uma diferença muito grande entre os gêneros nessa lista.

## 4.6. Quantidade Bilionários brasileiros divididos por categoria

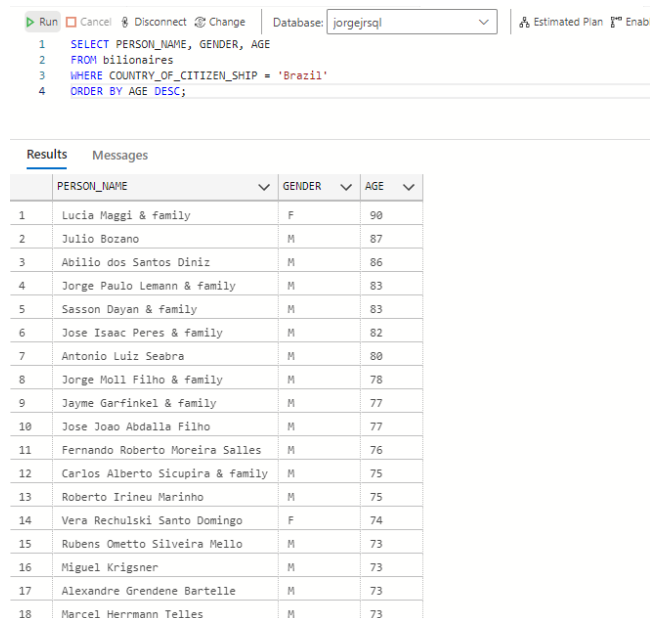


The screenshot shows a SQL query executed in a database client. The query is: `SELECT DISTINCT CATEGORY, COUNT(*) AS QUANTIDADE FROM billionaires WHERE COUNTRY_OF_CITIZENSHIP = 'Brazil' GROUP BY CATEGORY ORDER BY QUANTIDADE DESC;` The results are displayed in a table with two columns: CATEGORY and QUANTIDADE. The categories are ranked from highest to lowest count.

	CATEGORY	QUANTIDADE
1	Finance & Investments	12
2	Food & Beverage	10
3	Diversified	8
4	Fashion & Retail	6
5	Manufacturing	5
6	Media & Entertainment	3
7	Healthcare	3
8	Energy	2
9	Real Estate	1
10	Technology	1

As empresas financeiras estão no topo da categoria dos bilionários no Brasil, porém vemos o setor de alimentação na segunda colocação, cenário esse diferente do encontrado quando analisamos o ranking mundial.

## 4.7. Bilionários brasileiros por idade



The screenshot shows a SQL query executed in a database client. The query is: `SELECT PERSON_NAME, GENDER, AGE FROM billionaires WHERE COUNTRY_OF_CITIZENSHIP = 'Brazil' ORDER BY AGE DESC;` The results are displayed in a table with three columns: PERSON\_NAME, GENDER, and AGE. The results are ranked from highest to lowest age.

	PERSON_NAME	GENDER	AGE
1	Lucia Maggi & family	F	90
2	Julio Bozano	M	87
3	Abilio dos Santos Diniz	M	86
4	Jorge Paulo Lemann & family	M	83
5	Sasson Dayan & family	M	83
6	Jose Isaac Peres & family	M	82
7	Antonio Luiz Seabra	M	80
8	Jorge Moll Filho & family	M	78
9	Jayne Garfinkel & family	M	77
10	Jose Joao Abdalla Filho	M	77
11	Fernando Roberto Moreira Salles	M	76
12	Carlos Alberto Sicupira & family	M	75
13	Roberto Irineu Marinho	M	75
14	Vera Rechulski Santo Domingo	F	74
15	Rubens Ometto Silveira Mello	M	73
16	Miguel Krigsner	M	73
17	Alexandre Grendene Bartelle	M	73
18	Marcel Herrmann Telles	M	73

Em questão de idade, temos uma mulher como bilionária mais velha no Brasil.

## **5- Auto-análise**

Não tinha ideia do quanto eram interessantes esses ambientes de Cloud da Google, Amazon e Microsoft.

Resolvi iniciar pelo Google Cloud, porém me deparei com alguns problemas no Deploy no Data Fusion e não conseguia encontrar a solução. Deletei e criei todas as configurações algumas vezes, porém não consegui concluir com sucesso. Como solução resolvi fazer o projeto pelo AWS.

Cai de novo em alguns problemas, não consegui fazer com que o teste de conexão fosse concluído com sucesso. Criei em zonas distintas, mas não consegui subir a conexão. Foi então que migrei para o Microsoft Azure e a experiência foi das melhores. Então resolvi firmar nessa plataforma e com sucesso.

Quanto a análise de dados, acho que por todos esses problemas encontrados, não consegui utilizar o melhor do meu conhecimento em SQL, devendo algumas consultas e detalhes que esse relatório de bilionários nos permitia explorar. Valeu muito a experiência dessa primeira sprint, pois foi muito importante todo ensinamento teórico e prático passado.