

Análise de reputação com base em comentários no Twitter

Amanda Duarte Cordeiro, Astor Cesar de Souza Almeida, Deborah Cristina Hubinger, Frederico Esteves Martins, Jennyfer Oliveira da Silva

Orientador: Marco Paulo Soares Gomes

PUC Minas

Curso de Tecnologia em Banco de Dados

amandaduarte@pucminas.br, astormg@gmail.com,
deborahcristina1004@gmail.com, fred.critica@gmail.com,
jennyfersilva356@gmail.com

Resumo: As redes sociais são espaços na Internet que permitem a criação e compartilhamento de conteúdo, sendo fontes importantes de informações e opiniões. A indústria cinematográfica é uma das que se beneficia das redes sociais, pois são utilizadas como meios de divulgação de filmes e fontes de opiniões que podem influenciar e entender o comportamento dos usuários.

Pensando neste cenário e nos riscos e desafios de implementar novas regras, a empresa de streaming Netflix, encontrou a EstatBD uma companhia especializada em pesquisa, estatística, analytics e big data para atuar em um serviço de Web Scraping, coletando sentimentos expressivos dos usuários do Twitter, após o comunicado de repressão ao compartilhamento de senhas aos usuários que não residem no mesmo local.

Palavras-chave: análise de sentimentos; usuários; compartilhamento; Netflix; redes sociais; Twitter.

1. Introdução

Para atender às necessidades do mercado, as empresas têm apostado cada vez mais em Big Data e Analytics, onde é identificado a partir da gestão correta dos dados a oportunidade de maximizar os lucros, pensado em atender a essas necessidades foi criada a empresa de assessoria estatística, big data e analytics EstatBD SA.

A EstatBD SA oferece serviços especializados que tem como objetivo fornecer suporte em questões relacionadas à pesquisa, estatística, analytics e big data. Dessa forma, nossos serviços entregam o mapeamento de necessidades, planejamento de projetos e execução das estratégias com base em ferramentas tecnológicas e conceitos avançados em análise de dados.

Sendo formada por uma equipe de 5 funcionários especialistas de diversas áreas que envolvem machine learning, engenheiros de nuvem (Azure e AWS) e analistas de dados, reunimos competências que combinam com os recursos avançados de tecnologia e estratégias de Business Intelligence para resolução dos problemas.

Os serviços prestados pela assessoria EstatBD SA são diversos como, levantamento de requisitos, modelagem de banco de dados para atender a demandas analíticas, planejamento de projetos e suas execuções e análise de mercado para entender a visão do cliente de empresas diversas.

Atualmente, a EstatBD SA está prestando um serviço de Web Scraping (técnicas de coletas de dados de plataformas online para utilizar em análises futuras) para uma empresa de serviços de stream, a Netflix. Empresa cuja preocupação é a visão pública que possui, necessitando de feedbacks diversos dos clientes.

A Netflix estendeu sua repressão ao compartilhamento de senhas para o Canadá, Nova Zelândia, Portugal e Espanha. Os usuários desses países agora terão que pagar para dar às pessoas com quem não moram, acesso à sua conta, depois que regras semelhantes foram testadas na América Latina no ano passado. Os usuários desses países agora terão que pagar para dar às pessoas com quem não moram acesso à sua conta.

Essa medida vem sendo testada pela empresa em alguns países, e o que tudo indica é que ampliem as novas regras antes de março de 2023.

Estima-se que mais de 100 milhões de famílias compartilham suas contas. Por conta disso, uma nova página “gerenciar acesso e dispositivos” permitirá que os membros controlem mais facilmente quem tem acesso.

Em sua carta aos acionistas, a Netflix disse que, com base em sua experiência na América Latina, espera que uma proporção de usuários cancele suas assinaturas assim que as mudanças sejam implementadas, mas prevê que o número geral de usuários crescerá com o tempo.

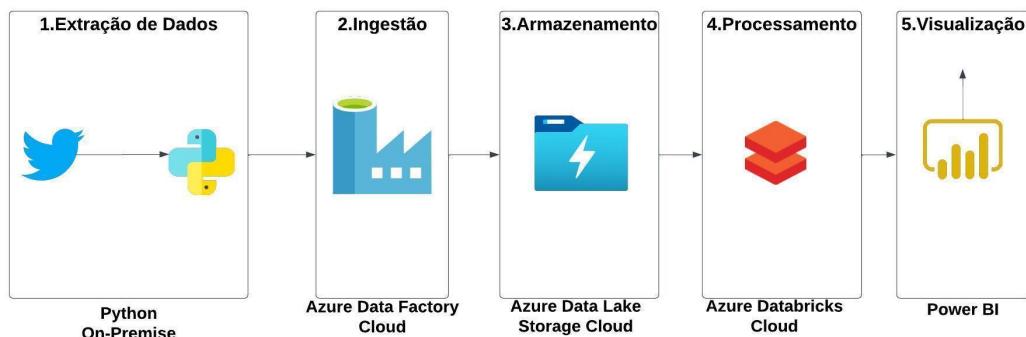
A empresa teve uma queda de 50% em suas ações no ano passado devido a preocupações com as assinaturas e suas empresas concorrentes, mas as ações tiveram um aumento de 24% desde o início do ano, impulsionadas pela queda do dólar americano.

Portanto, ao se deparar com os riscos e desafios de se implementar uma nova regra, a empresa encarou a análise e os estudos dos dados que vão sendo gerados por internautas, como uma necessidade. Dados estes que podem ser encontrados em forma de “feedbacks” em plataformas de comunicação em tempo real, como o Twitter.

2.Arquitetura

A arquitetura do projeto foi elaborada em 5 etapas conforme figura e descrição a seguir:

Figura 1: Arquitetura do projeto



Fonte: Os Autores (2023)

Etapa 1: Extração dos dados web no Twitter utilizando técnica de WebScraping executada na ferramenta Jupyter Notebook em computador On-Premises.

Etapa 2: Ingestão dos dados brutos com origem On-Premises para destino em Cloud utilizando Azure Data Factory.

Etapa 3: Armazenamento do arquivo bruto na ferramenta de Data Lake Azure Data Lake Storage e transformação do arquivo utilizando Python.

Etapa 4: Processamento in cloud do arquivo bruto limpando os dados e transformando utilizando o recurso Azure Databricks com processamento em cluster.

Etapa 5: Visualização do resultado da etapa anterior com Microsoft Power BI.

3. Metodologia

Para coletar os dados do twitter, sobre as opiniões dos usuários da plataforma de streaming Netflix, utilizaremos o processo de extração web Scraping, que é muito utilizado em estratégias de transformação digital, bem como para automatizar processos de coleta e consulta de dados.

Inicialmente pensamos na utilização da API (Interface de Programação de Aplicações), como os dados do Twitter são comumente usados para fins de marketing e pesquisa, a documentação da API possui limitações de acesso e um alto custo de utilização, por este motivo, optamos pelo desenvolvimento do web scraper que permite a criação da nossa própria API, que pode ser vinculada a qualquer site ou plataforma.

Para realização desta extração utilizaremos a aplicação Jupyter Notebook, para a criação da web scraping, que fará a raspagem dos dados que desejamos coletar da rede social Twitter.

3.1 Web Scraping

Foi desenvolvido, na IDE do Jupyter Notebook, um código em Python para a realização da raspagem de dados, metodologia também conhecida como Web Scraping. Com isso, será possível extrair Twitters da página oficial do Twitter com a hashtag #NetflixBrasil, com datas de corte de 03/09/2022 até 03/01/2023, tendo em vista que o plano de assinatura básico da Netflix foi implantado em 03/11/2022 foi definido dois meses depois e antes dessa data para avaliar os Twitters. Segue o código em Python que foi utilizado:

Figura 2: Arquivo bruto extraído

```
date,url,username,content
2023-01-02 22:32:17+
00:00,https://twitter.com/hospicionerd/status/1610041314388721
667,hospicionerd,"Ótimo sucesso que chama?
AlchemyOfSouls novamente entre os assuntos mais comentados do
Twitter. Os novos episódios estarão disponíveis na
#NetflixBrasil a partir de 21 de janeiro.
#APNerd https://t.co/kG11AtLYmN"
2023-01-02 21:48:13+
00:00,https://twitter.com/jonathan_cabral/status/1610030227618
467842,jonathan_cabral,A Netflix faliu quando cancelou essa
aqui #Netflix #NetflixBrasil https://t.co/BbGqmzF97y
2023-01-02 20:13:40+
00:00,https://twitter.com/mayafalcao_/status/16100064333528023
07,mayafalcao_,"Na real?
Não gostei de Olhar Indiscreto...
#NetflixBrasil https://t.co/juWXn8HpLz"
2023-01-02 20:03:19+
00:00,https://twitter.com/Bia_Snape/status/1610003828752977921
,Bia_Snape,"Eu não acredito que a Netflix cancelou #1899
Que vergonha, enquanto isso investem nessas séries porcarias
que a história só ""mais do mesmo""
#Netflix #netflixbrasil #1899Netflix #Dark #darkserie #
1899series"
2023-01-02 16:17:18+
00:00,https://twitter.com/TKelesidan/status/160994694763212800
1,TKelesidan,#Netflixbrasil os episódios 16 à 22 da nova
temporada nos Power Rangers estão em outro idioma...
2023-01-02 14:23:42+
00:00,https://twitter.com/chandellyinfos/status/16099183593818
76738,chandellyinfos,"Ótimo hortolani via Instagram:
(Re) FRESCANDO Ótimo
#ocangaceirodofuturo #netflixbrasil https://t.co/r50GLI8Mif"
2023-01-02 12:57:26
```

Fonte: Os Autores (2023)

####Atualização do pip se necessário#####

```

pip install --upgrade pip

####Instalação da biblioteca que raspda dados do Twitter####

!pip install snscreape

```

```

import snscreape.modules.twitter as sntwitter
import pandas as pd
import os

#Define diretório local
os.chdir("D:/PUC BD/P4/Projeto")

#Faz a raspagem dos Twitters até um Limite máximo definido
maxTweets = 100000000
i = 0
tweets_list = []
for tweet in sntwitter.TwitterSearchScraper('#NetflixBrasil since:2022-09-03
until:2023-01-03').get_items():
    if i > maxTweets:
        break
    tweets_list.append([tweet.date, tweet.url, tweet.user.username, tweet.rawCo
ntent])
    i = i + 1

#Transforma o resultado da raspagem dos dados em data frame
tweets_df = pd.DataFrame(tweets_list, columns=['date', 'url', 'username', 'con
tent'])

#Exporta data frame como .CSV para diretório Local
tweets_df.to_csv('Tweet.csv', encoding = 'utf-8', index = False)

```

3.2 Criar grupo de recurso no Azure

Seguem os passos e os prints de da criação de um grupo de recursos no Azure para fazer uma melhor gestão dos recursos que serão criados:

Entre no portal Azure e clique em mais serviços:

The screenshot shows the Microsoft Azure homepage. At the top, there's a banner with three cards: 'Começar com uma avaliação gratuita do Azure', 'Gerenciar o Azure Active Directory', and 'Acessar os benefícios do aluno'. Below this, there's a section titled 'Serviços do Azure' with various service icons. A red box highlights the 'Mais serviços' button at the bottom right of this section. Further down, there's a 'Recursos' section with tabs for 'Recente' (selected) and 'Favorito', and a 'Última visualização' section.

Depois clique em Grupos de recursos:

The screenshot shows the 'Todos os serviços | Todo' page in the Microsoft Azure portal. On the left, there's a sidebar with various service categories. The main area shows a grid of service icons. In the 'Geral' section, the 'Grupos de recursos' item is highlighted with a red box. Other items in this section include 'Todos os recursos', 'Assinaturas', 'Marketplace', 'Modelos', 'Centro de Início Rápido', and 'Reservas'.

Criar:

Página inicial > Grupos de recursos > [https://portal.azure.com/#blade/HubsBlade/resourceType/groups/resourceName/](#)

Grupos de recursos ...

[+ Criar](#) [Gerenciar a exibição](#) [Atualizar](#) [Exportar para CSV](#) [Abrir a consulta](#) [Atribuir marcas](#)

Filtros: Criar | [Filtrar](#) Criar | [Filtrar](#) Assinatura igual a tudo | [Localização igual a tudo](#) | [Adicionar filtro](#)

Nome: Assinatura: Localização:

0 Recursos não seguros | 0 Recomendações

Não há grupos de recursos para exibir.
Experimente alterar ou limpar seus filtros.

[Criar grupo de recursos](#) | [Saiba mais](#) | [?](#)

[Enviar comentários](#)

Defina um nome para o seu grupo de recursos:

Página inicial > Grupos de recursos > [https://portal.azure.com/#blade/HubsBlade/resourceType/groups/resourceName/](#)

Criar um grupo de recursos ...

Básico | Marcações | Revisar + criar

Grupo de recursos: Um container que armazena recursos relacionados a uma solução do Azure. O grupo de recursos pode incluir todos os recursos para a solução ou apenas os recursos que você deseja gerenciar como um grupo. Você decide como deseja alocar recursos para os grupos de recursos com base no que é mais conveniente para sua organização. [Saiba mais](#)?

Detalhes do projeto

Assinatura: Grupo de recursos:

Detalhes do recurso

Região:

[Revisar + criar](#) | [Avançar: Marcações >](#)

Avance até Revisar + Criar e clique em criar:

Screenshot of the Microsoft Azure 'Criar um grupo de recursos' (Create a resource group) page. The validation status is 'Validação aprovada.' (Validation approved). The basic configuration shows the following details:

- Assinatura (Subscription): Azure for Students
- Grupo de recursos (Resource Group): Projeto_Big_Data
- Região (Region): East US

The 'Marcações' (Tags) section shows 'Nenhum' (None). At the bottom, there are buttons for 'Criar' (Create), 'Anterior' (Previous), 'Avançar' (Next), and 'Baixar um modelo para automação' (Download a template for automation).

Pronto agora o grupo de recursos está criado.

Screenshot of the Microsoft Azure 'Grupos de recursos' (Resource Groups) page. The 'Projeto_Big_Data' resource group is listed in the results. The 'Nome' (Name) column shows the resource group names, and the 'Projeto_Big_Data' row is highlighted with a red box.

3.3 Criação de recursos (Azure Data Factory, Azure Data Lake Storage Gen2 e Azure Databricks)

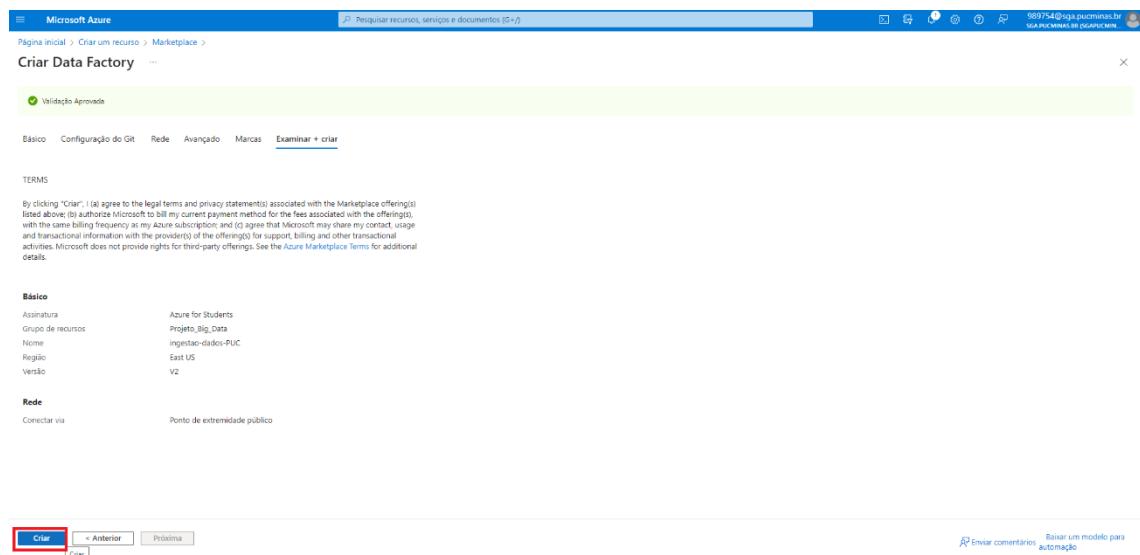
3.3.1 Azure Data Factory

Utilizaremos o Azure Data Factory para a ingestão dos dados. O Data Factory permite o ETL (extração, transformação e carregamento), depois que os dados forem transformados e carregados no armazenamento. A transformação dos dados será tratada com atividades nativas e instrumentos do Data Factory como o fluxo de dados. Os dados serão processados com um código Python personalizado.

Volte à página inicial, clique em criar um recurso e filtre usando o termo “Data factory” finalizando com “Criar”:

Selecione o grupo de recursos já criado e defina um nome para o Data Factory:

Clique em avançar mantendo as configurações padrão até Examinar + Criar, clique em Criar:



3.3.2 Azure Data Lake Storage Gen2

Os dados extraídos através do processo de web scraping, serão armazenados no Azure Data Lake que inclui todos os recursos necessários para que seja mais fácil armazenar dados e realizar todo tipo de processamento e análise. Com ele removeremos as complexidades relacionadas com a ingestão o armazenamento dos dados, proporcionando escala e desempenho na nuvem. Com o Azure Data Lake poderemos analisar todos os seus dados em um único lugar, sem restrições.

No campo de procura, digite “Armazenamento” e clique em “Contas de armazenamento”:

Clique em criar e em seguida selecione o grupo de recursos criados, nome da conta de armazenamento e redundância LRS finalizando com Avançar:

Selecione a caixa do Data Lake Storage Gen 2 e avançar:

Microsoft Azure

Página inicial > Contas de armazenamento >

Criar uma conta de armazenamento

Básico Avançado Rede Proteção de dados Criptografia Marcas Examinar + criar

Segurança

Defina as configurações de segurança que afetam sua conta de armazenamento.

- Exigir transferência segura para operações da API REST ⓘ
- Habilitar o acesso público ao blob ⓘ
- Habilitar o acesso à chave de conta de armazenamento ⓘ
- Padrão para autorização do Azure Active Directory no portal do Azure ⓘ
- Versão mínima do TLS ⓘ Versão 1.2
- Escopo permitido para operações de cópia (versão prévia) ⓘ De qualquer conta de armazenamento

Data Lake Storage Gen2

O namespace hierárquico do Data Lake Storage Gen2 acelera cargas de trabalho de análise de big data e habilita ACLs (listas de controle de acesso) de nível de arquivo. [Saiba mais](#)

Habilitar namespace hierárquico



Armazenamento de Blobs

Review

< Anterior

Avançar: Rede >

Clique em avançar mantendo o padrão até Examinar + Criar espere ser examinado e clique em criar para criar seu Lake Storage:

Criar uma conta de armazenamento

Básico Avançado Rede Proteção de dados Criptografia Marcas Examinar + criar

Conectividade de rede Ponto de extremidade público (todas as redes)
Nível de roteamento padrão Roteamento de rede da Microsoft
Tipo de ponto de extremidade Padrão

Proteção de dados

Restauração pontual Desabilitado
Exclusão reversível do blob Habilidado
Período de retenção de blob em dias 7
Exclusão reversível do contêiner Habilidado
Período de retenção do contêiner em dias 7
Exclusão reversível do compartilhamento Habilidado
de arquivo
Período de retenção do compartilhamento 7
de arquivo em dias
Controle de versão Desabilitado
Feed de alterações do blob Desabilitado
Suporte à imutabilidade no nível da versão Desabilitado

Criptografia

Tipo de criptografia MMK (chaves gerenciadas pela Microsoft)
Habilitar suporte para chaves gerenciadas pelo cliente Sómente blobs e arquivos
Habilitar a criptografia de infraestrutura Desabilitado

Criar

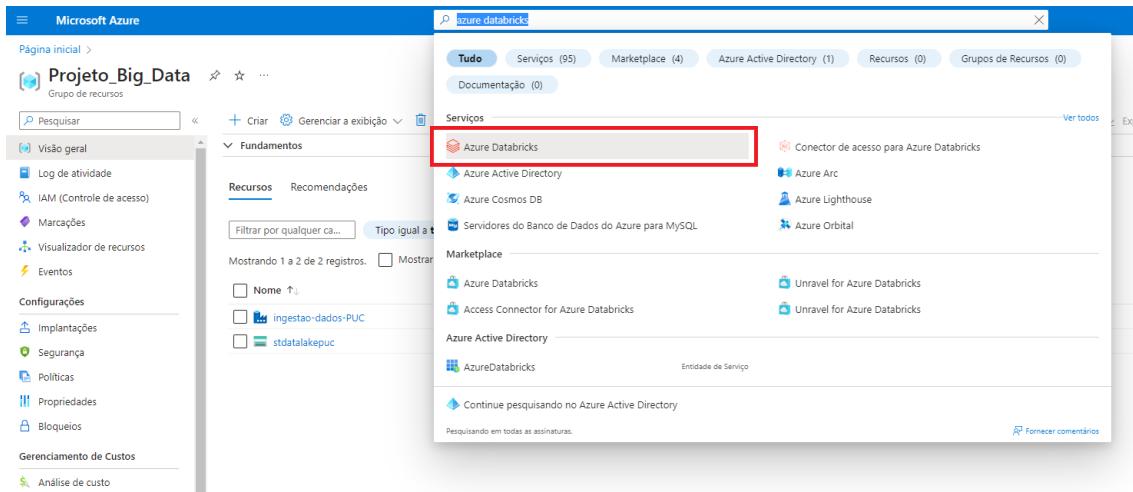
< Anterior

Avançar > Baixar um modelo para automação

3.3.3 Azure Databricks

O recurso Azure Databricks será criado para processar os dados armazenados no Data Lake de forma rápida e eficiente.

Na barra de procura digite “Azure Databricks”, clique nele e finalize com “Criar”:



Em Noções básicas, selecione o grupo de recursos criado, defina o nome para o recurso e o tipo de preço pode ficar como Standard:

Criar um workspace do Azure Databricks

Noções Básicas Em Rede Criptografia Marcas Revisar + criar

Detalhes do Projeto

Selezione a assinatura para gerenciar os custos e os recursos implantados. Use grupos de recursos como pastas para organizar e gerenciar todos os seus recursos.

Assinatura * Grupo de recursos * Nome do Workspace * Região * Tipo de Preço *

Standard (Apache Spark, Seguro com o Azure AD)
Standard (Apache Spark, Seguro com o Azure AD)
Premium (+ Controles de acesso baseados em função)
Avaliação (Premium – DBUs Gratuitos por 14 Dias)

Em Rede, Criptografia, Marcas, Revisar + cria pode deixar padrão e clicar em “Criar”. Pronto, recurso criado.

3.4 Criar conexão entre o Azure Data Factory e Azure Data Lake Storage Gen2

Na página inicial clique no seu grupo de recursos:

Serviços do Azure

+ Criar um recurso Azure Databricks Contas de armazenamento Grupos de recursos Educação Serviços de Aplicativos Centro de Início Rápido Máquinas virtuais Bancos de dados SQL Mais serviços

Recursos

Recente Favorito

Nome	Tipo	Última visualização
Projeto_Big_Data	Grupo de recursos	8 minutos atrás
Processamento-DF	Serviço do Azure Databricks	10 minutos atrás
statalakepuc	Conta de armazenamento	uma hora atrás

Navegar

Assinaturas Grupos de recursos Todos os recursos Painel

Ferramentas

Microsoft Learn Azure Monitor Microsoft Defender para Nuvem Gerenciamento de Custos

Selecione o Azure Data Factory:

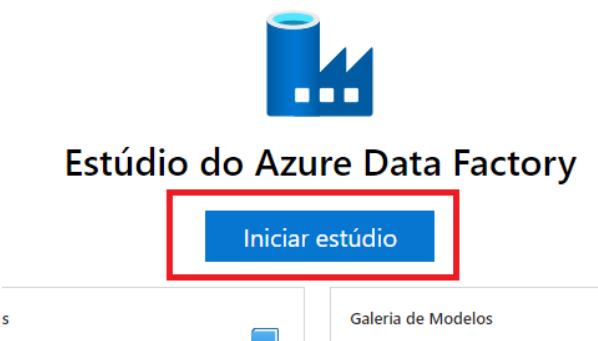
Recursos Recomendações

Filtrar por qualquer ca... Tipo igual a tudo Localização igual a tudo Adicionar filtro

Mostrando 1 a 3 de 3 registros. Mostrar os tipos ocultos Nenhum agrupamento

Nome	Tipo	Localização
ingestao-dados-PUC	Data factory (V2)	East US
ingestao-dados-PUC	Serviço do Azure Databricks	East US
ingestao-dados-PUC	Conta de armazenamento	East US

Clique em Iniciar estúdio para abrir o recurso:



Clique em manager e Linked Services:

The screenshot shows the Azure Data Factory 'Manage' section. The left sidebar has buttons for Home, Author, Monitor, Manage (which is highlighted with a red box), and Learning Center. The main content area shows a 'General' section with a 'Configurações do data fact...' link, followed by a 'Connections' section with a 'Linked services' link (also highlighted with a red box). The 'Linked services' section displays a note: 'O serviço vinculado define as informações de conexão para um armazenamento de dados ou uma computação.' and a 'Novo' button. Below this is a table with columns 'Nome', 'Tipo', and 'Relacionado'. At the bottom, it says 'Nenhum serviço vinculado a ser mostrado' and 'Criar serviço vinculado'.

Clique em Novo, digite na barra de procura data lake e selecione Azure Data Lake Storage Gen2:

Crie um nome para a conexão, selecione a assinatura Azure, nome da conta de armazenamento e finalize clicando em criar:

Novo serviço vinculado

Azure Data Lake Storage Gen2 [Saiba mais](#)

Nome *
AzureDataLakeStorage1

Descrição

Conectar-se via runtime de integração *
AutoResolveIntegrationRuntime

Tipo de Autenticação
Chave de conta

Método de seleção de conta
 Da assinatura do Azure Inserir manualmente

Assinatura do Azure
Azure for Students (6410eba2-f013-49e5-8d2b-25aed89d7c90)

Nome da conta de armazenamento *
stdatalakepuc

Testar conexão
 Para o serviço vinculado Para o caminho do arquivo

Anotações
+ Novo

> Parâmetros
> Avançado

Criar Voltar Testar conexão Cancelar

Agora vamos criar um container no data lake para receber os dados do Data Factory, na página inicial ou no grupo de recursos selecione a conta de armazenamento:

Recursos

Recente Favorito

Nome	Tipo	Última visualização
stdatalakepuc	Conta de armazenamento	4 minutos atrás
ingestao-dados	Data factory (V2)	18 minutos atrás
Projeto_Big_Dat	Grupo de recursos	18 minutos atrás
Processamento	Serviço do Azure Databricks	30 minutos atrás

Exibir Excluir

Ver todos

Clique em contêineres:

Página inicial > stdatalakepuc

stdatalakepuc | Contêineres

Contá de armazenamento

Pesquisar

+ Contêiner Alterar o nível de acesso

Eventos

Navegador de armazenamento (versão prévia)

Armazenamento de dados

Contêineres

Compartimentos de arquivos

Filas

Tabelas

Segurança + rede

Rede

Chaves de acesso

Assinatura de acesso compartilhado

Criptografia

Pesquisar contêineres por prefixo

Nome

\$logs

Clique em + Contêiner, defina o nome e o nível de acesso ao container:

Novo contêiner

Nome *
stg

Nível de acesso público
Contêiner (acesso de leitura anônimo para contêineres e blo...)

Todos os clientes do blob e do contêiner podem ver blobs por solicitação anônima. Os clientes podem enumerar blobs dentro do contêiner por solicitação anônima, mas não podem enumerar os contêineres dentro da conta de armazenamento. O acesso anônimo ignora as configurações de ACL (Lista de Controle de Acesso).

Avançado

Criar Give feedback ↗

Criar conexão com máquina local, no azure da factory clique em Manage, Integration runtimes e Novo:

Runtimes de integração

O IR (runtime de integração) é a infraestrutura de computação para fornecer as funcionalidades.

Novo

Nome	Tipo	Subtipo
AutoResolveIntegrationR...	Azure	Público
conect-local	Auto-hospedado	---

Selecione Auto-hospedado e continuar:

Configuração do runtime de integração

Ambiente de rede:

Escolha o ambiente de rede da fonte de dados/destino ou a computação externa com a qual o runtime de integração se conectará para fluxos de dados ou atividades de expedição ou movimentação de dados:



Azure

Use esta opção para executar fluxos de dados, movimentação de dados, atividades externas e de pipeline em uma computação totalmente gerenciada e sem servidor no Azure.



Auto-hospedado

Usar para executar atividades em uma rede local/privada

[Exibir mais](#) ▾

Recursos Externos:

Você pode usar um runtime de integração auto-hospedada que existe em outro recurso. Desse modo, é possível reutilizar a infraestrutura existente na qual o runtime de integração auto-hospedada está configurado.



Auto-Hospedado Vinculado

[Saiba mais](#) ▾

[Continuar](#)

[Voltar](#)

[Cancelar](#)

Defina o nome da integração e clique em criar. Baixe o programa da opção 2 e instale na sua máquina local:

Editar runtime de integração

Configurações Nós Atualização automática Compartilhamento Links

Instale o runtime de integração no computador Windows ou adicione mais nós usando a Chave de Autenticação.

Nome

Descrição

Opção 1: configuração expressa
[Clique aqui para iniciar a configuração expressa para este computador](#)

Opcão 2: configuração manual

Etapa 1: Baixar e instalar o runtime de integração

Etapa 2: usar essa chave para registrar o seu runtime de integração

Nome Chave de autenticação

Chave1  

Chave2  

Aplicar **Cancelar**

Após instalar o programa na máquina local, insira no campo aberto a chave 1 do passo anterior e clique em registrar:

Microsoft Integration Runtime Configuration Manager

Register Integration Runtime (Self-hosted)

Welcome to Microsoft Integration Runtime Configuration Manager. Before you start, register your Integration Runtime (Self-hosted) node using a valid Authentication Key.



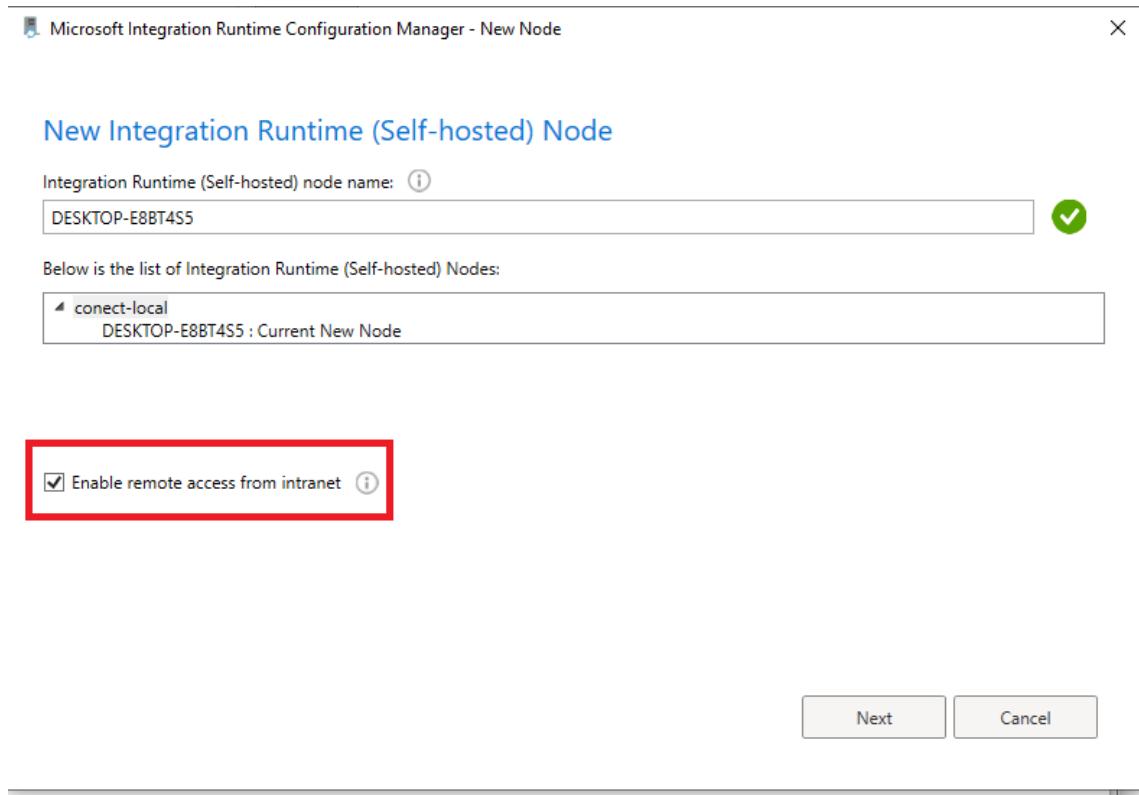
Show Authentication Key [Learn how to find the Authentication Key](#)

HTTP Proxy
Current Proxy: No proxy **Change**

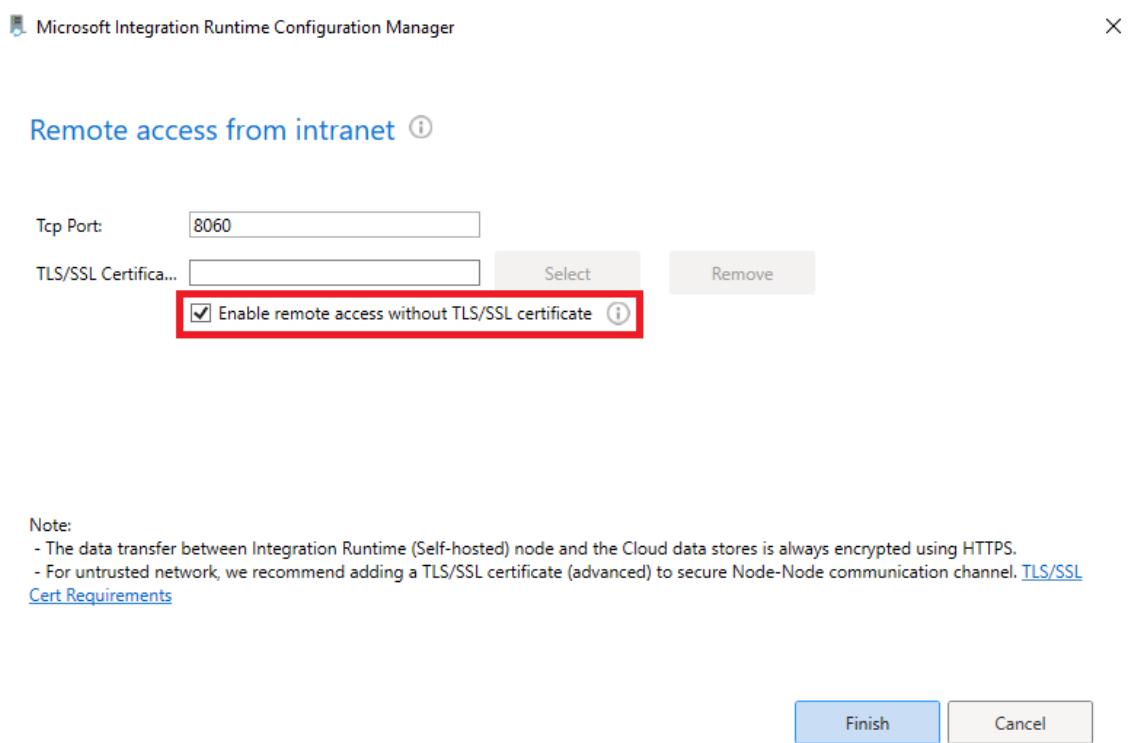
Diagnostic Tool
[Troubleshoot problems \(preview\)](#)

Register **Cancel**

Selezione “Enable remote access from intranet” e clique em Next:



Selecione “Enable remote access without TLS/SSL certificate” e clique em “Finish”:



Clique em “Launch Configuration Manager”:

 Microsoft Integration Runtime Configuration Manager



Register Integration Runtime (Self-hosted)

Welcome to Microsoft Integration Runtime Configuration Manager. Before you start, register your Integration Runtime (Self-hosted) node using a valid Authentication Key.

Authentication Key 

Show Authentication Key [Learn how to find the Authentication Key](#)

 Integration Runtime (Self-hosted) node has been registered successfully.

Note: You can associate up to 4 physical nodes with a Self-hosted Integration Runtime. This enables high availability and scalability for the Self-hosted Integration Runtime.
We recommend you setup at least 2 nodes for higher availability. [See Integration Runtime \(Self-hosted\) article for details.](#)

HTTP Proxy

Current Proxy: No proxy 

Diagnostic Tool

[Troubleshoot problems \(preview\)](#)  

E pronto, foi criado um nó entre a máquina local e a cloud da Azure:

 Microsoft Integration Runtime Configuration Manager

 [Home](#) [Settings](#) [Diagnostics](#) [Update](#) [Help](#)

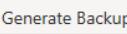
 Self-hosted node is connected to the cloud service

Data Factory: ingestao-dados-PUC
Integration Runtime: conect-local
Node: DESKTOP-E8BT4S5



Data Source Credential

Credential store: On-premises
Credential status: In sync
Last backup time: N/A

 Connected to the cloud service (Data Factory V2) 

3.5 Ingestão dos dados utilizando Azure Data Factory

Processo de ingestão de arquivos CSV local pelo Data Factory com fluxo para armazenamento no Data Lake Azure.

Clique no Azure Data Factory, Manage, Linked services e Novo:

The screenshot shows the Microsoft Azure Data Factory interface. On the left, there's a sidebar with options: Home, Author, Monitor, **Manage** (which is highlighted with a red box), and Learning Center. Under 'Manage', there are links for General, Configurações do data fact..., Connections, Integration runtimes, Microsoft Purview, Source control, Git configuration, ARM template, Author, and Triggers. The main content area is titled 'Serviços vinculados' (Linked services). It says 'O serviço vinculado define as informações de conexão para um armazenamento de dados ou uma computação.' (A linked service defines connection information for a data storage or computation). There's a '+ Novo' (New) button with a red box around it. Below it is a search bar 'Filtrar por nome' (Filter by name) and a note 'Anotações: Qualquer um'. A table lists one item: 'Nome' (Name) is 'AzureDataLakeStorage1', 'Tipo' (Type) is 'Azure Data Lake Storage Gen2', and 'Relacionad' (Related) has a value of '0'. At the bottom, there are buttons for 'Validar tudo' (Validate all) and 'Publicar tudo' (Publish all).

Procure arquivo e selecione sistemas de arquivo:

The screenshot shows the 'Novo serviço vinculado' (New linked service) dialog. At the top, there are tabs for 'Armazenamento de dados' (selected) and 'Computação'. A search bar contains the text 'arquivo'. Below the search bar is a navigation bar with buttons for 'Tudo', 'Arquivo', 'Azure', 'Banco de dados', 'NoSQL', and 'Protocolo genérico'. Two items are listed in a grid: 'Armazenamento de Arquivos do Azure' (with a blue hexagonal icon) and 'Sistema de arquivos' (with a grey server icon). Both items are highlighted with red boxes.

Defina o nome, integração local, Host(Diretório local), usuário do computador local e senha.

Clique em testar e depois em salvar:

Editar serviço vinculado

Sistema de arquivos [Saiba mais](#)

Nome *
localcsv

Descrição

Conectar-se via runtime de integração * ⓘ
 conect-local

Host * ⓘ
D:

Nome do usuário *
Frederico

Senha [Azure Key Vault](#)

Senha *

Anotações

+ Novo

> Parâmetros

> Avançado ⓘ

[Salvar](#) [Cancelar](#) [Testar conexão](#)

Criar pipeline de dados para ingestão:

Microsoft Azure | Data Factory > ingestao-dados-PUC

Data Factory | Validar tudo | Publicar tudo 1

Recursos do Data Factory

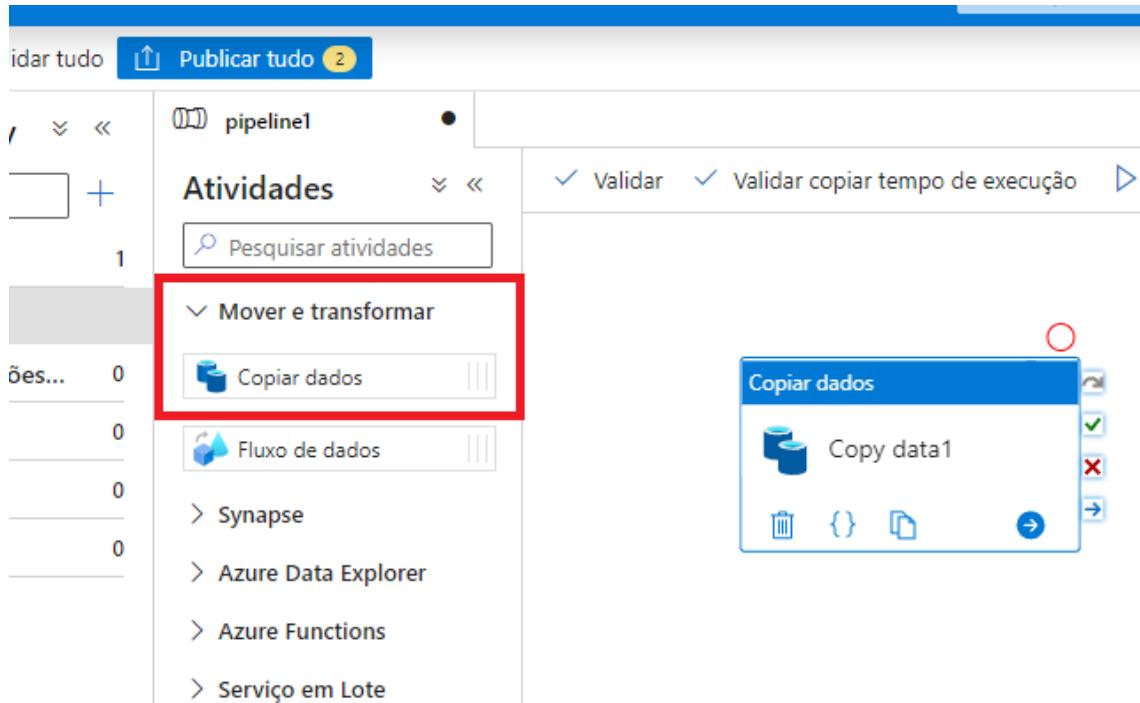
Filtrar os recursos por nome +

Pipelines 0

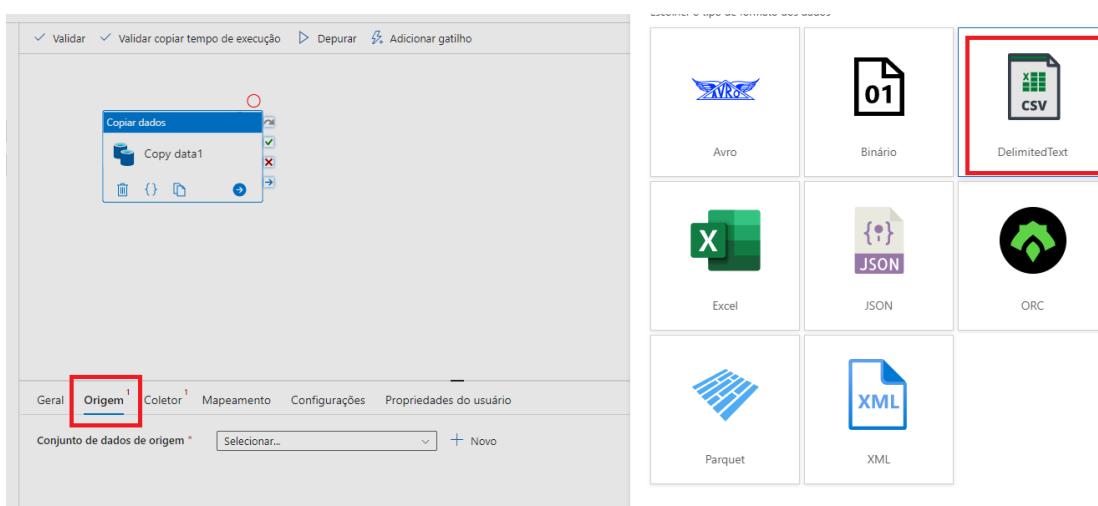
Novo pipeline

- Captura de Dados de Alterações...
- Datasets
- Data flows
- Power Query

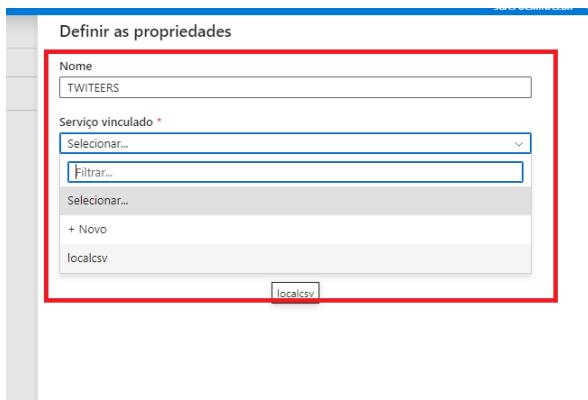
Em atividade clique em Mover e transformar, clique e arraste o objeto “Copiar dados” para o workflow:



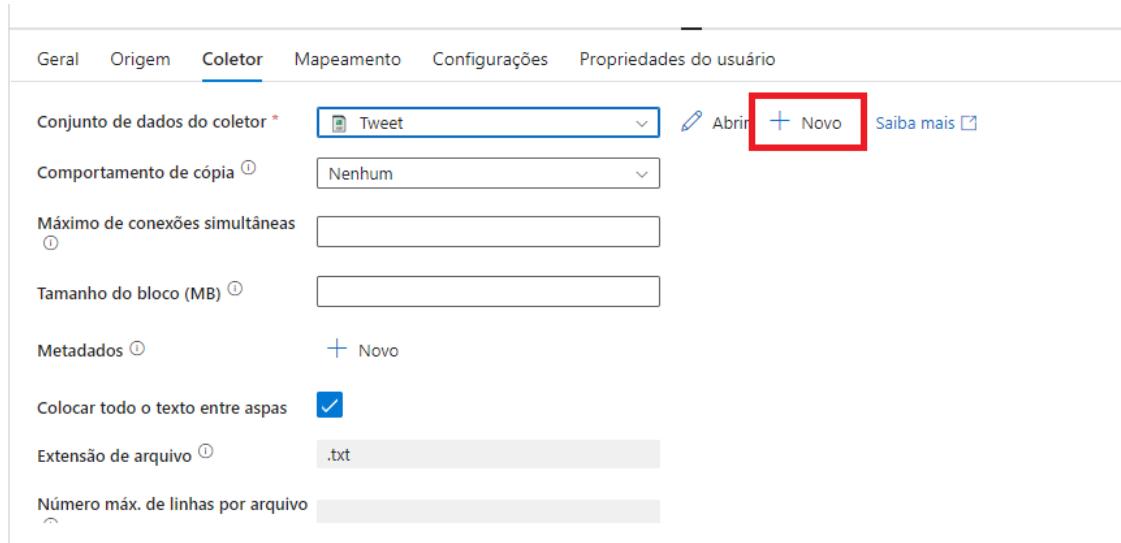
Clique em copiar dados e na barra inferior onde está localizada a configuração do objeto “copiar dados” e selecione “Origem” para configurar a origem dos dados. Selecione o tipo .CSV:



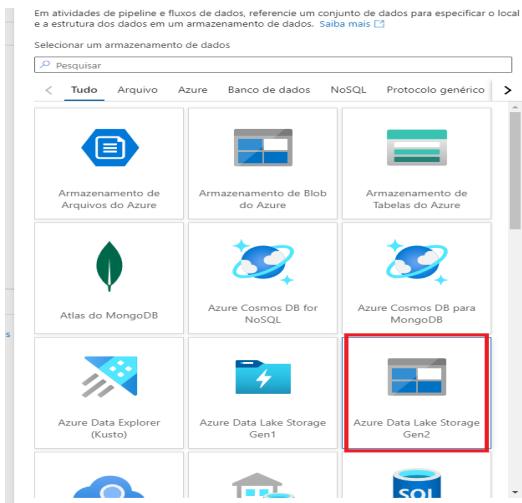
Selecione o nome da origem, serviço local criado, valide o diretório e clique em ok:



Definindo o destino, em “Coletor” clique em Novo:



Selecione o Azure Data Lake Storage Gen 2:



Selecione o tipo do arquivo, que nesse caso é .csv, defina o diretório do container no Storage e, por fim, clique em Validar tudo e depois depurar:

Nome	Tipo	Inicio da execução	Duração	Status	Runtime de integração	ID da exec.
Copy data1	Copiar dados	2023-03-26T02:55:16.1	00:00:32	Bem-sucedido	connect-local	efdc5bec

Pronto o processo de ingestão foi efetuado carregando um arquivo bruto tipo CSV da máquina local para um Data Lake do Azure através do Azure Data Factory.

Link que baixa o arquivo do container:

https://stdatalakepuc.blob.core.windows.net/stg/Projeto_PUC/Tweet.csv

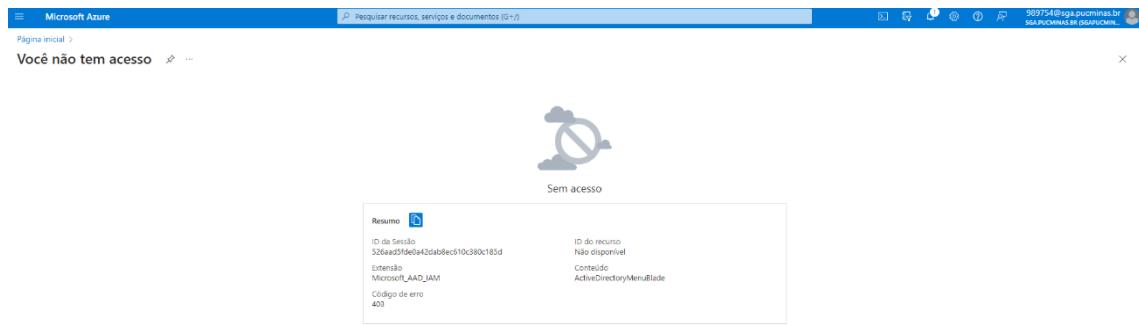
3.6 Limpeza e transformação do arquivo bruto carregado no Data Lake

Infelizmente não foi possível conectar o Azure Databricks ao Data Lake para a transformação do arquivo bruto ingerido, pois a licença concedida pela PUC Minas para a utilização dos recursos da Azure possui uma limitação imposta pelo administrador de TI da mesma, segue evidências:

The screenshot shows the 'Compute / Frederico Martins's Cluster' configuration page in the Azure Databricks interface. Under the 'Advanced options' section, there is a red box highlighting the 'Azure Data Lake Storage credential passthrough' field. This field is described as being available on Azure Databricks premium and links to documentation for 'Using Data Lake Storage Gen2 with Azure Databricks'. Below this, there are sections for 'Spark config' (with an example of spark.speculation: true) and 'Environment variables' (containing PYSPARK_PYTHON=/databricks/python/bin/python3).

Para que a conexão entre o Databricks e o Data Lake seja estabelecida é necessário utilizar o recurso Azure Active Directory, porém não temos acesso:

The screenshot shows the 'Serviços do Azure' (Azure Services) blade in the Azure portal. A red box highlights the 'Criar um recurso' (Create a resource) section, which contains a card for 'Azure Active Directory'. Below this, the 'Recursos' (Resources) section shows a list of recent resources, with 'Recente' (Recent) selected. The 'Azure Active Directory' card is also visible in the 'Exibir' (View) section.



Sendo assim, a etapa de transformação e novo carregamento dos dados foram feitos utilizando código Python. Segue o código Python utilizado, print do resultado e do carregamento, o carregamento foi feito utilizando a etapa 3.5.

```
import os
import pandas as pd

#Define diretório local
os.chdir("D:/PUC BD/P4/Projeto")
tweets_df.to_csv('Tweet.csv', encoding = 'utf-8', index = False)

#!pip install nltk
import nltk

import random

#!pip install unidecode

from unidecode import unidecode
from nltk.corpus import stopwords

from nltk.metrics import *
import pandas as pd

# tokenizar / quebrar em termos / bag of words
from nltk.tokenize import word_tokenize
#nltk.download('punkt') # faz o download do "tokenizador" que será usado
em word_tokenize

# lista de stopwords
#nltk.download('stopwords')
```

```
from nltk.corpus import stopwords

# Remoção de acentuação
#!pip install unidecode
from unidecode import unidecode

# Stopwords
#nltk.download('stopwords')
stop_words = stopwords.words('portuguese')

tweets = tweets_df['content']

tweets_transformados = pd.DataFrame()

contador = 0

while contador < tweets.count():
    tokens = word_tokenize(tweets[contador])

    # transforma todas as palavras em minusculas
    palavras_minusculas = [tokens.lower() for tokens in tokens]

    # Exclui espaços a mais
    exclui_espacos = [palavras_minusculas.strip() for palavras_minusculas in palavras_minusculas]

    # Remove caracteres especiais
    exclui_esp = [palavras_minusculas for palavras_minusculas in palavras_minusculas if palavras_minusculas.isalnum()]

    # Remove acentuacoes
    remov_acent = [unidecode(exclui_esp) for exclui_esp in exclui_esp]

    # Remove conectivos
    stop_words = stopwords.words('portuguese')
    remov_conect = [remov_acent for remov_acent in remov_acent if not remov_acent in stop_words]

    # Remove -2 letras para excluir pq, q , oc etc
    remov_abrev = [remov_conect for remov_conect in remov_conect if len(remov_conect) >= 2]
```

```
contador = contador + 1

print(remov_abrev)

tweets_transformados = tweets_transformados.append(remov_abrev)

tweets_transformados.rename(columns={0:
'Tweets'}).to_csv('Tweet_transformados.csv', encoding='utf-8',
index=False)
```

Evidência do arquivo resultante em .csv:

	A
1	Tweets
2	sucesso
3	chama
4	alchemyofsouls
5	novamente
6	assuntos
7	comentados
8	twitter
9	novos
10	episodios
11	estarao
12	disponiveis
13	netflixbrasil
14	partir
15	21
16	janeiro
17	apnerd
18	https
19	netflix
20	faliu
21	cancelou
22	aqui
23	netflix
24	netflixbrasil
25	https
26	real
27	nao
28	gostei

Tweet_transformados

Pronto  Acessibilidade não disponível

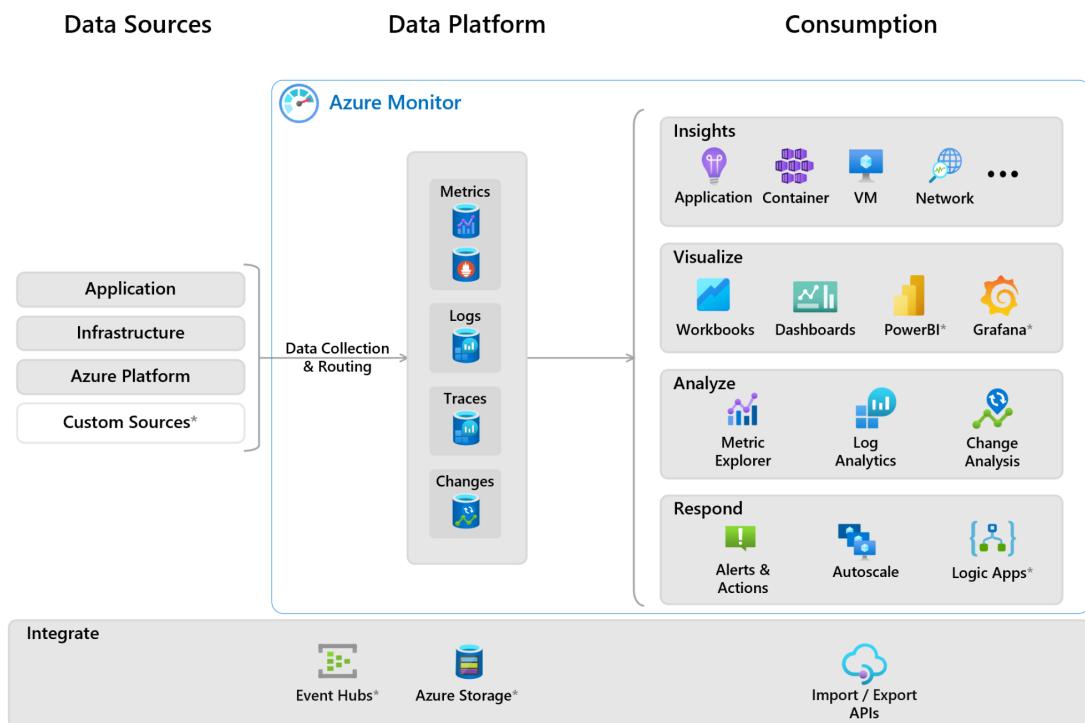
Evidência do arquivo no Data Lake:

Azure Storage Container blade showing the contents of a blob container named 'stg'. The table lists three files: 'I_J.csv', 'Tweet_transformados.csv', and 'Tweet.csv'. The first two files are highlighted with a red box.

Nome	Modificado	Camada de acesso	Status do arquivo	Tipo de blob	Tamanho	Estado de concessão
I_J.csv	22/04/2023, 16:04:30	Principal (inferidos)		Blob de blocos	191.37 KB	Disponível
Tweet_transformados.csv	25/03/2023, 23:55:45	Principal (inferidos)		Blob de blocos	381.54 KB	Concedido
Tweet.csv						

4. Recurso do Azure para observabilidade: Monitor Azure

Para permitir a observabilidade, os dados devem ser coletados de todas as camadas e componentes do sistema distribuído. Esta plataforma faz a coleta e a agregação de dados de diversas fontes em uma plataforma onde poderá ser utilizado para análise, visualização e alerta.



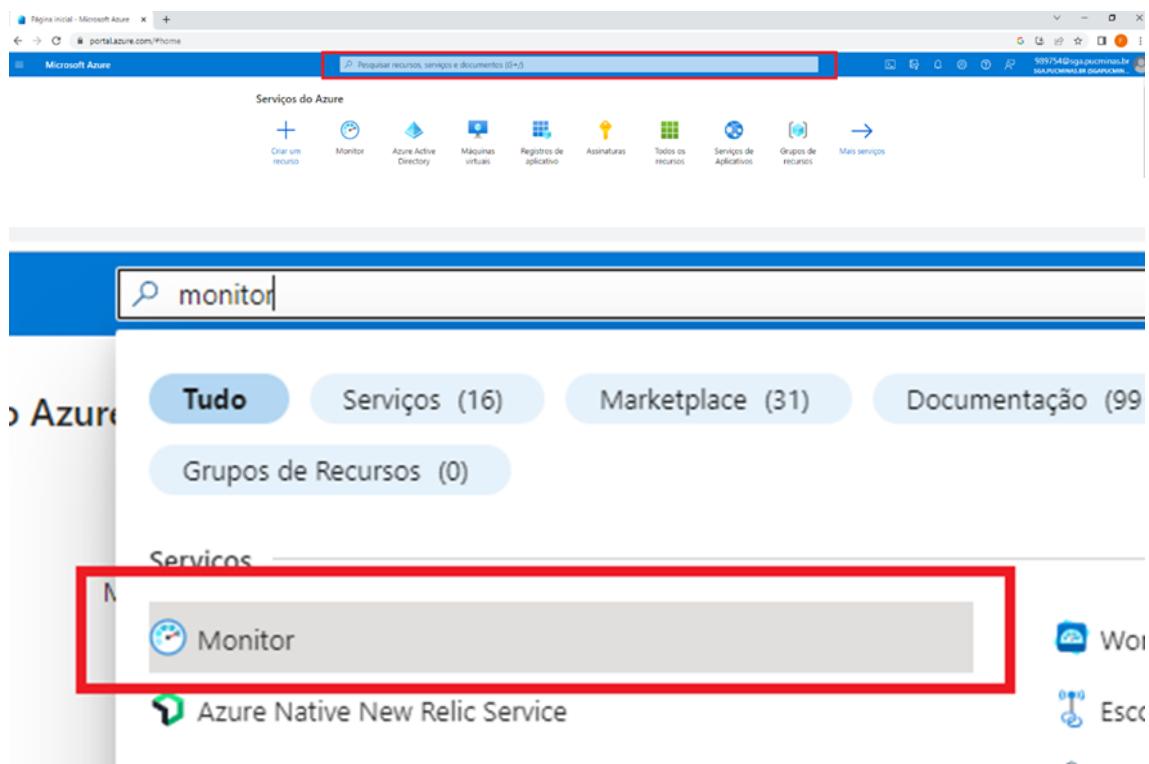
Fonte: Microsoft Learn Azure Monitor

A EstatBD SA com a necessidade de observar métricas, logs e rastreamento distribuídos escolheu a plataforma do Azure Monitor para suprir a

demandas. A ferramenta consegue coletar esses três tipos diferentes de dados para fornecer observabilidade suficiente de um sistema monitorado.

Os recursos do Azure são otimizados e geram quantidades significativas de dados de monitoramento suficientes para o projeto que a EstatBD está atuando, consolidando dados de diversas fontes em uma plataforma de métricas ou logs, sendo interessante para acompanhar as métricas de armazenamento e processamento de forma a manter o controle de gastos e possíveis erros que possam consumir processamento e armazenamento desnecessários.

Para utilizar o recurso do Monitor Azure basta digitar monitor no buscador do portal azure e selecionar Monitor:



Escolha o tipo de monitoramento, no nosso caso vamos utilizar o monitoramento por métricas e alertas:

Página inicial >

Monitor | Visão Geral

Microsoft

Pesquisar

Visão Geral

Log de atividades

Alertas

Métricas

Logs

Análise de Alterações

Integridade do Serviço

Pastas de trabalho

Insights

Aplicativos

Máquinas Virtuais

Contas de armazenamento

Contêineres

Redes

SQL (versão prévia)

Visão Geral Tutoriais Insights

Use as exibições de monitoramento selecionadas para recursos específicos do Azure. Exibir todos os insights

Application Insights

Monitore a disponibilidade, o desempenho, os erros e o uso do seu aplicativo.

Exibir Mais

Insights do Contêiner

Obtenha a visibilidade do desempenho e da integridade de seus controladores, nós e contêineres.

Exibir Mais

Detecção, triagem e diagnóstico

Visualize, analise e responda aos dados e eventos do monitoramento. Saiba mais sobre o monitoramento

Métricas

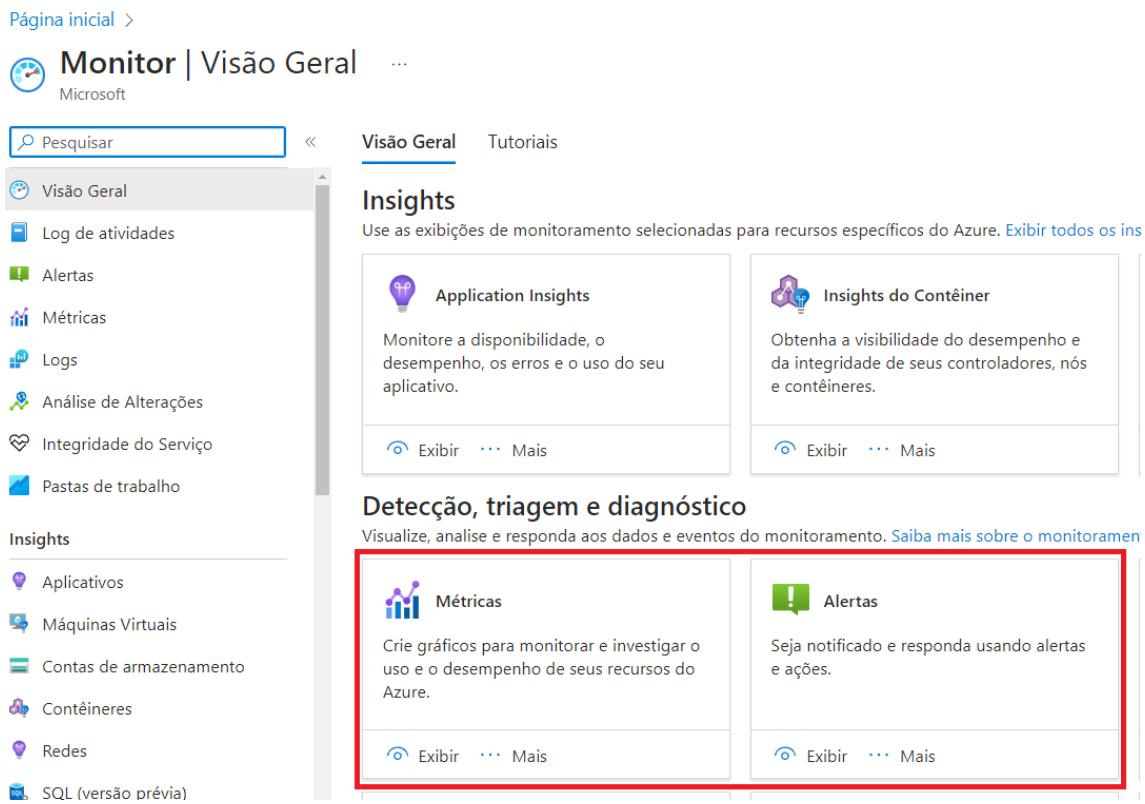
Crie gráficos para monitorar e investigar o uso e o desempenho de seus recursos do Azure.

Exibir Mais

Alertas

Seja notificado e responda usando alertas e ações.

Exibir Mais



Configuração do monitor por métricas, clique em Exibir, selecione o escopo e aplicar:

Página inicial > Monitor

Monitor | Métricas

Microsoft

Pesquisar

Visão Geral

Log de atividades

Alertas

Métricas

Logs

Análise de Alterações

Integridade do Serviço

Pastas de trabalho

Insights

Aplicativos

Máquinas Virtuais

Contas de armazenamento

Contêineres

Redes

SQL (versão prévia)

Azure Cosmos DB

Cofres de Chaves

Novo gráfico Atualizar Compartilhar Comentários

Título do gráfico

Escopo Selecionar um escopo Namespace de Métrica Selecionar namespace

100
90
80
70
60
50
40
30
20
10
0

Wendy

Selecionar um escopo

Procurar Recente

Tipos de recurso

Todos os tipos de recurso

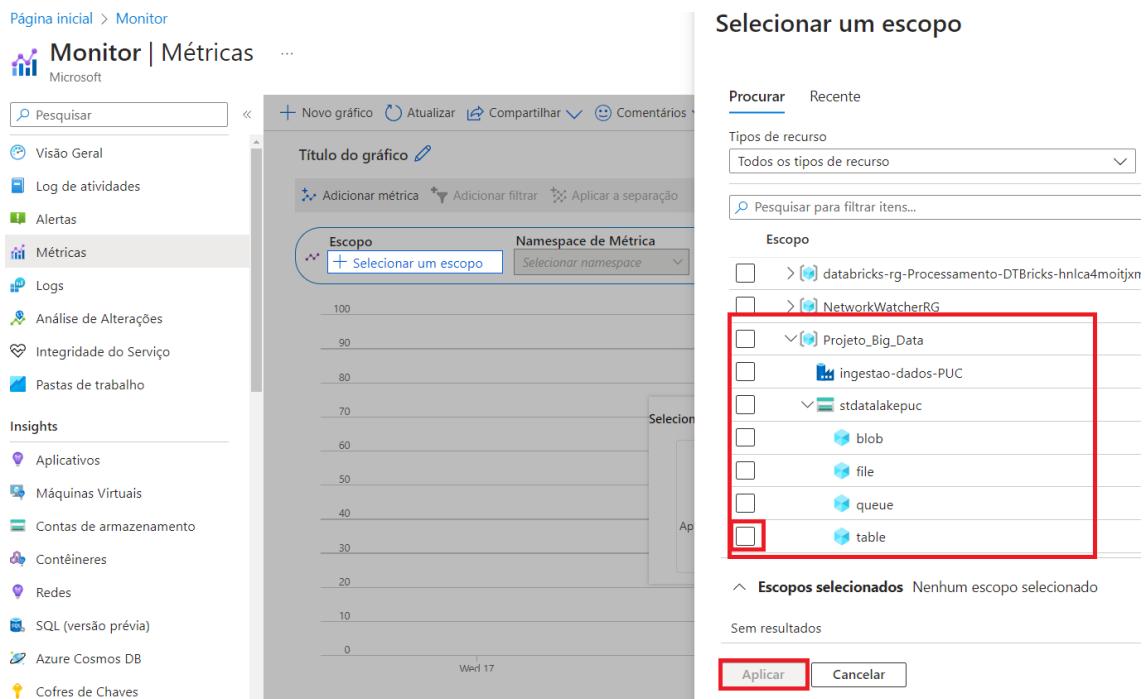
Pesquisar para filtrar itens...

Escopo

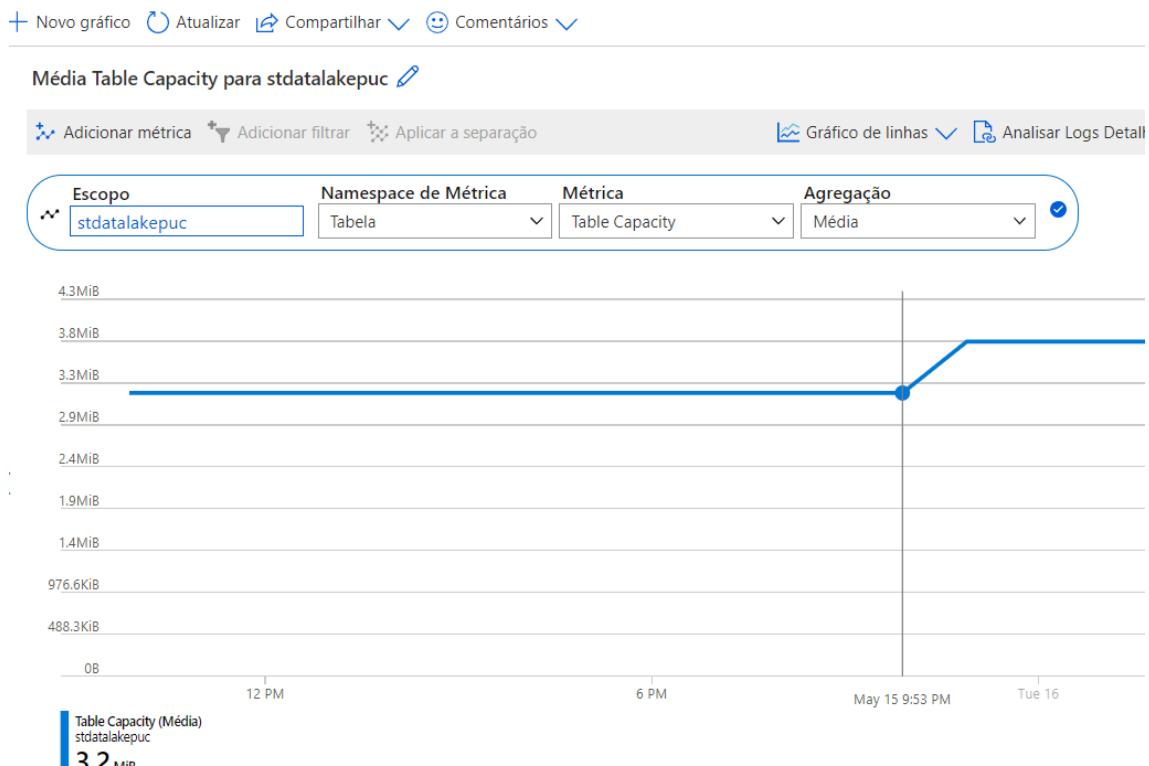
Escopos selecionados Nenhum escopo selecionado

Sem resultados

Aplicar Cancelar



Escolha a métrica a ser monitorada, vamos selecionar Table Capacity para monitorar a capacidade de armazenamento:

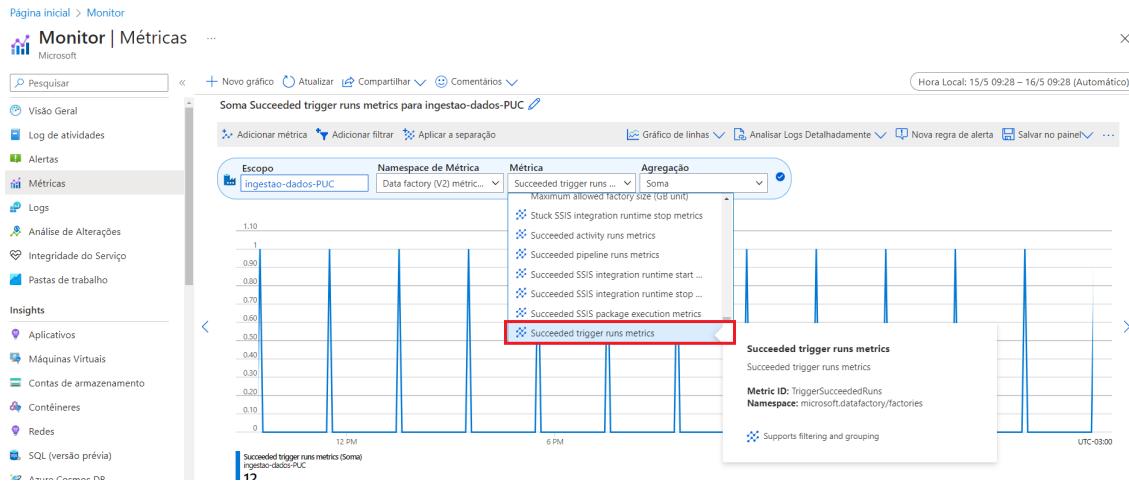


No dia 15 de maio é possível visualizar no gráfico uma elevação, esse período foi um momento em que adicionamos um arquivo para teste no Data Lake apenas para demonstrar no gráfico essas flutuações, tendo em vista que o arquivo objeto deste estudo é estático e é demonstrado na constante da reta.

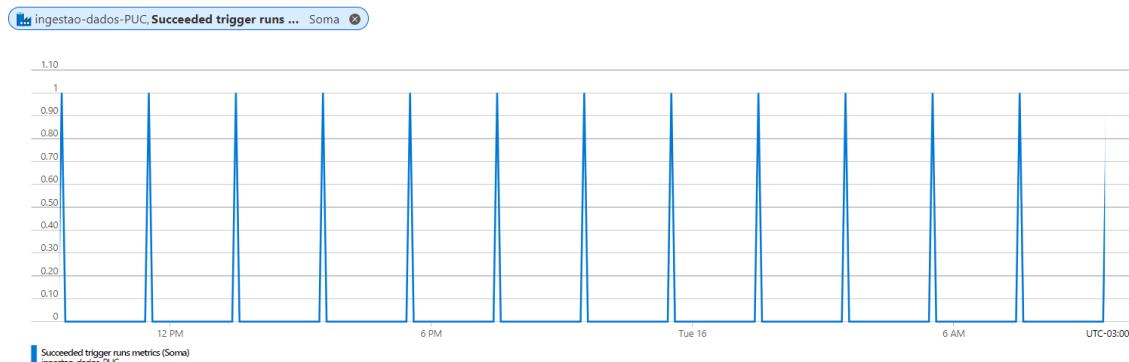
Uma outra métrica interessante é utilizar o escopo do Data Factory para monitorar o sucesso das cargas agendadas:

The screenshot shows the Azure Metrics Monitor interface. On the left, there's a sidebar with navigation links like 'Visão Geral', 'Log de atividades', 'Alertas', 'Métricas' (which is selected), 'Logs', 'Análise de Alterações', 'Integridade do Serviço', 'Pastas de trabalho', 'Insights', 'Aplicativos', 'Máquinas Virtuais', 'Contas de armazenamento', 'Contêineres', 'Redes', 'SQL (versão prévia)', 'Azure Cosmos DB', and 'Cofres de Chaves'. The main area displays a chart titled "Média Table Capacity para stdatalakepuc". To the right, a modal window titled "Selecionar um escopo" (Select scope) is open, listing resources. The table in the modal has columns for "Escopo" (Scope), "Tipo de recurso" (Resource type), and "Localização" (Location). One row is highlighted with a red box: "ingestao-dados-PUC" (selected), "Data factory (V2)", and "East US". Other rows include "stdatalakepuc" (Assinatura, East US) and "blob" (microsoft.storage/storageaccounts, East US). A note at the bottom of the modal says: "Por que não consigo selecionar vários recursos? A seleção múltipla com métricas não foi habilitada para Data factory (V2) recursos. Você pode informar à equipe Data factory (V2) que essa funcionalidade é importante e votar nessa solicitação." (Why can't I select multiple resources? Multi-selection with metrics is not enabled for Data factory (V2) resources. You can report to the Data factory (V2) team that this functionality is important and vote on this request.)

Selecione a métrica Succeeded pipeline run metrics:



Perceba que a cada duas horas ocorre um pico no gráfico, esses picos representam o sucesso da execução do pipeline de dados criado no Data Factory:



Por fim, monitoramento por alerta, clique no monitor de alerta, criar e selecione Regra de alerta:

The screenshot shows the Microsoft Azure Monitor Alerts dashboard. The top navigation bar includes 'Microsoft Azure' and a search bar. The main area displays a summary of alert rules: 0 errors, 0 warnings, 0 informational, and 0 detailed. A large green circle with a checkmark indicates 'Nenhum alerta disparado no 24 horas atrás' (No alerts triggered in the last 24 hours). The left sidebar lists various monitoring categories like 'Visão Geral', 'Log de atividades', 'Alertas', 'Métricas', 'Logs', 'Análise de Alterações', 'Integridade do Serviço', and 'Pastas de trabalho'. The 'Alertas' section is currently selected. The bottom status bar shows the date and time as '17/05/2023 21:48'.

Vou selecionar o escopo do Data Lake:

The screenshot shows the 'Criar uma regra de alerta' (Create alert rule) wizard. The current step is 'Selecionar um recurso' (Select resource). On the left, there's a sidebar with tabs for 'Escopo', 'Condição', 'Ações', 'Detalhes', 'Marcas', and 'Examinar + criar'. The 'Escopo' tab is selected. Below it, there's a note: 'Crie uma regra de alerta para identificar e tratar problemas quando forem encontradas condições importantes nos recursos.' Under 'Recurso', it says 'Ainda não há nenhum recurso selecionado'. On the right, the 'Selecionar um recurso' interface shows a list of resources under 'Recursos'. One item, 'stdatalakepuc', is selected and highlighted with a red border. Other items include 'Azure for Students', 'CEAL_TESTE_ASTOR', 'databricks-rg-Processamento-DTBricks-hnlica4moijxm', 'NetworkWatcherRG', 'Projeto_Big_Data', 'ingestao-dados-PUC', 'Processamento-DTBricks', and 'stdatalakepuc'. Below the list, a section titled 'Recursos selecionados' shows '1 conta de armazenamento' selected: 'stdatalakepuc' (Conta de armazena... East US). At the bottom are 'Aplicar' and 'Cancelar' buttons.

Vou selecionar a condição Used capacity para criar um alerta quando meu Data Lake chegar a um nível de armazenamento limite:

Microsoft Azure

Página inicial > Monitor | Alertas >

Criar uma regra de alerta ...

Escopo **Condição** Ações Detalhes Marcas Examinar + criar

Configure quando a regra de alerta deve ser disparada selecionando um sinal e definindo a lógica.

Nome do sinal *

Used capacity

Popular

Availability

Used capacity

Success E2E Latency

Transactions

Success Server Latency

Pesquisa de logs personalizada

[See all signals](#)

Lógica de alerta

Limite

Tipo de agregação

Operador

Unidade

Valor do limite *

Obrigatório
 O limite deve ser um número positivo

[Examinar + criar](#)

[Anterior](#)

[Avançar: Ações >](#)

Foi definido o operador maior que, unidade em MB, valor limite 200mb, e frequência de verificação a cada 1 hora:

Microsoft Azure 989754@sga.pucminas.br SGA.PUCMINAS.BR (SGAPUCMIN..)

Página inicial > Monitor | Alertas >

Criar uma regra de alerta ...

[See all signals](#)

Lógica de alerta

Limite
 Estático Dinâmico
Os Limites Dinâmicos não estão disponíveis para esta métrica no momento

Tipo de agregação
 Média

Operador
 Maior que

Unidade
 MiB

Valor do limite *

Quando avaliar

Verificar a cada

Período de retroativo

[+ Adicionar condição](#)

Visualização 0.10 USD/mês

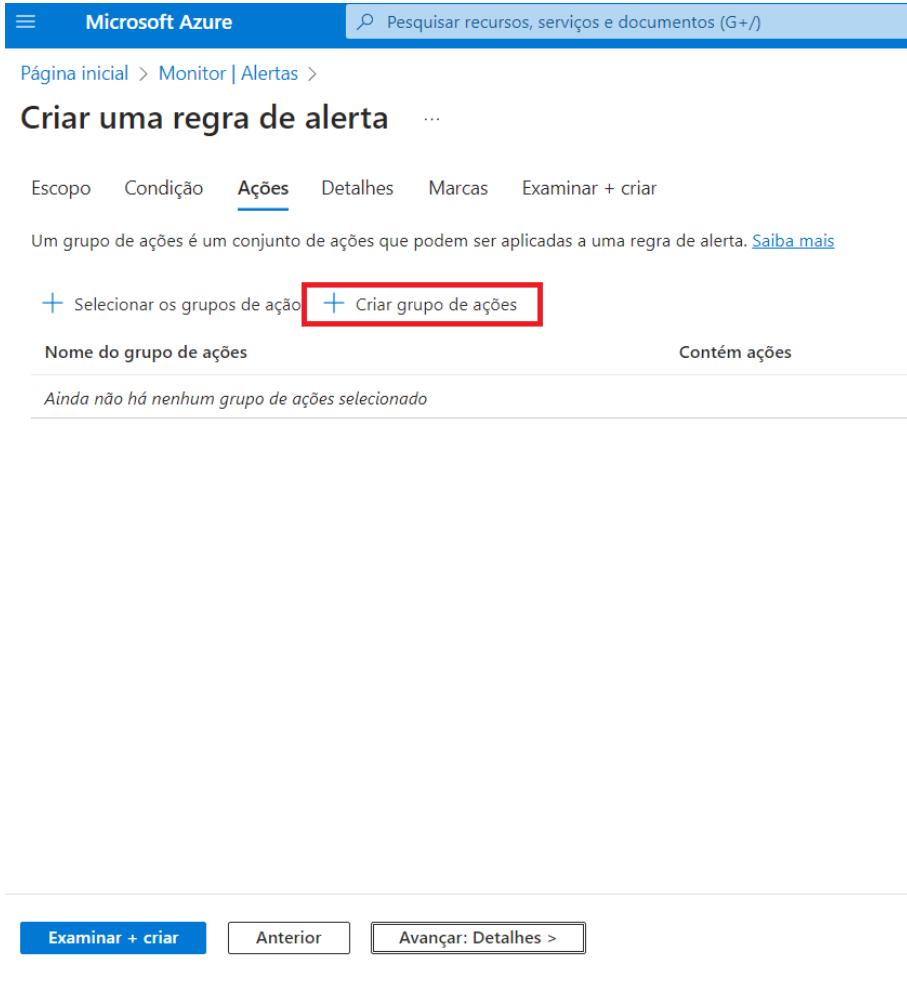
Sempre que Used capacity average for maior que 200 Mebibyte

Intervalo de tempo : Nas últimas 6 horas Série temporal : Agregação

Used capacity (Media)
4,35 MB

[Examinar + criar](#) [Anterior](#) [Avançar: Ações >](#)

Na próxima etapa podemos definir uma ação, caso não tenha nenhuma criada é necessário à sua criação:



Página inicial > Monitor | Alertas >

Criar uma regra de alerta

...
Escopo Condição Ações Detalhes Marcas Examinar + criar

Um grupo de ações é um conjunto de ações que podem ser aplicadas a uma regra de alerta. [Saiba mais](#)

+ Selecionar os grupos de ação + Criar grupo de ações

Nome do grupo de ações	Contém ações
Ainda não há nenhum grupo de ações selecionado	

Examinar + criar Anterior Avançar: Detalhes >

Defina o grupo de recursos e nome do grupo de ações:

Criar grupo de ações

Básico Notificações Ações Rótulos Examinar + criar

Um grupo de ações invoca um conjunto definido de notificações e ações quando um alerta é disparado. [Saiba mais](#)

Detalhes do projeto

Selecione uma assinatura para gerenciar custos e recursos implantados. Use grupos de recursos como pastas para organizar e gerenciar todos os recursos.

Assinatura *	Azure for Students
Grupo de recursos *	Projeto_Big_Data
Região *	Global

Detalhes da instância

Nome do grupo de ações *	Limite DLK
Nome de exibição *	Limite DLK

O nome de exibição é limitado a 12 caracteres

Examinar + criar **Anterior** **Próximo: Notificações >**

Notificação por Email:

Criar grupo de ações

Básico Notificações Ações Rótulos Examinar + criar

Escolha como será notificado quando o grupo de ações for disparado. Essa etapa é opcional.

Notificação tipo	Nome	Selecionado
Email/Mensagem SMS/Enviar por ...		

Email/Mensagem SMS/Enviar por Push/Ser...

Adicionar ou editar Email/Mensagem SMS/Enviar por Push/Serviço de Voz ação

<input checked="" type="checkbox"/> Email	Email * <input type="text" value="989754@sga.pucminas.br"/>
<input type="checkbox"/> SMS (Podem ser aplicados encargos da operadora)	Código do país <input type="text" value="1"/>
<input type="checkbox"/> aplicativo móvel do Azure notificação	Email da conta do Azure <input type="text"/>
<input type="checkbox"/> Serviço de Voz	Código do país <input type="text" value="1"/>

Habilite o esquema de alerta comum. [Saiba mais](#)

OK

Examinar + criar **Anterior** **Próximo: Ações >**

A ação é opcional, defina o nome para o Rótulo e clique em criar:

Página inicial > Monitor | Alertas > Criar uma regra de alerta >

Criar grupo de ações

Básico Notificações Ações Rótulos Examinar + criar

Este é um resumo do seu grupo de ações. Examine as informações para garantir que estejam corretas e considere o [Preços de Monitoramento do Azure](#) e a [Declaração de privacidade do Azure](#).

Básico

Assinatura	Azure for Students
Grupo de recursos	Projeto_Big_Data
Região	global
Nome do grupo de ações	Limite DLK
Nome de exibição	Limite DLK

Notificações

Tipo de notificação	Nome	Selecionado
Email/Mensagem SMS/Enviar por Push/Serviço de Voz	Fred	Email

Ações

Nenhum

Rótulos

Criar

[Anterior](#)

[Criar](#)

Grupo de ação criado, agora vamos para detalhes, nessa etapa podemos definir a gravidade do alerta, após a definição avance para marcas:

Página inicial > Monitor | Alertas >

Criar uma regra de alerta

Escopo Condição Ações **Detalhes** Marcas Examinar + criar

Detalhes do projeto

Selecione a assinatura e o grupo de recursos para salvar a regra de alerta.

Assinatura * ⓘ Azure for Students

Grupo de recursos * ⓘ Projeto_Big_Data

[Criar novo](#)

Detalhes da regra de alerta

Gravidade * ⓘ 3 - Informativo

Nome da regra de alerta * ⓘ

Descrição da regra de alerta ⓘ

Opções avançadas

3 - Informativo

10 - Crítico

1 - Erro

2 - Aviso

3 - Informativo

4 - Detalhado

Examinar + criar Anterior **Anterior: Marcas >**

Em marcas pode ser definido um grupo (opcional) e depois avance para a última etapa para criar:

Página inicial > Monitor | Alertas >

Criar uma regra de alerta

Escopo Condição Ações Detalhes Marcas **Examinar + criar**

Regra do alerta de métrica Preço total

1 Condição Variável

[Termos de uso](#) | [Política de privacidade](#) Preços

Escopo

Recurso Azure for Students > Projeto_Big_Data > stdatalakepuc

Condição

Nome do sinal	UsedCapacity
Operador	Maior que
Tipo de agregação	Média
Valor do limite	209715200
Período de retroativo	1 hora
Verificar a cada	1 minuto

Ações

Nome do grupo de ações	Contém ações
Criar	
Anterior	

O alerta criado fica ativado 24 horas, clicando em Regras de alerta é possível verificar o alerta criado e seu status:

Nome ↑	Condição	Severidade ↑↓	Escopo de destino	Tipo de recurso de destino	Tipo de sinal ↑↓	Status ↑↓
Regra estouro de limite DK	UsedCapacity > 209715200	0 - Crítico	stdatalakepuc	Conta de armazenamento	Métricas	Habilitado

Em resumo foi implantado o recurso de monitoramento Azure para monitorar a capacidade de armazenamento do Data Lake, accuracy de execução de pipeline de dados do Azure Data Factory e foi criado um alerta de limite de armazenamento no Data Lake.

5. Análise e Visualização de dados

No mundo atual em que vivemos, a quantidade de dados gerada diariamente é colossal. A capacidade de extrair informações significativas se tornou essencial para as empresas e estudantes de diversas áreas.

Essa análise envolve a coleta, organização, limpeza e transformação de dados brutos em informações comprehensíveis, nos permitindo encontrar padrões ocultos, insights valiosos e responder a perguntas importantes relacionadas ao assunto em questão do projeto, tais como:

1. A quantidade de usuários tweetando sobre a mudança do serviço da Netflix;
2. A quantidade de tweets a respeito do assunto;
3. Período.

5.1 Justificativa da ferramenta de visualização

Esta etapa foi decisiva para definir quais seriam as ferramentas de visualização de dados. Essa definição é importante pois dita como os usuários,

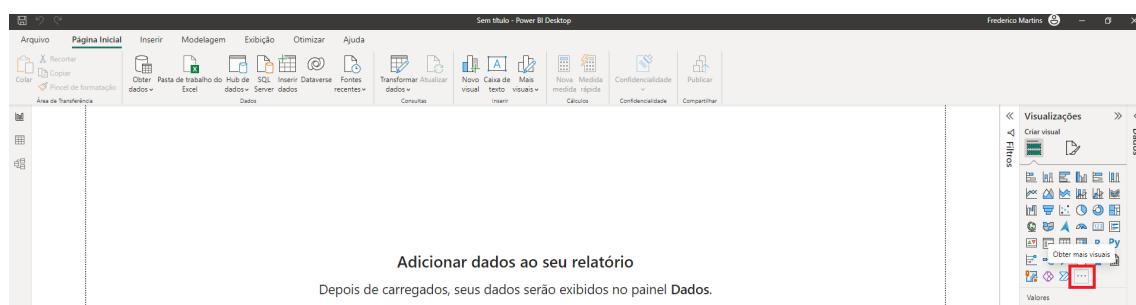
sejam da área de TI, analistas de dados da área de negócio ou da ponta do negócio, vão visualizar tanto os insights gerados quanto os indicadores.

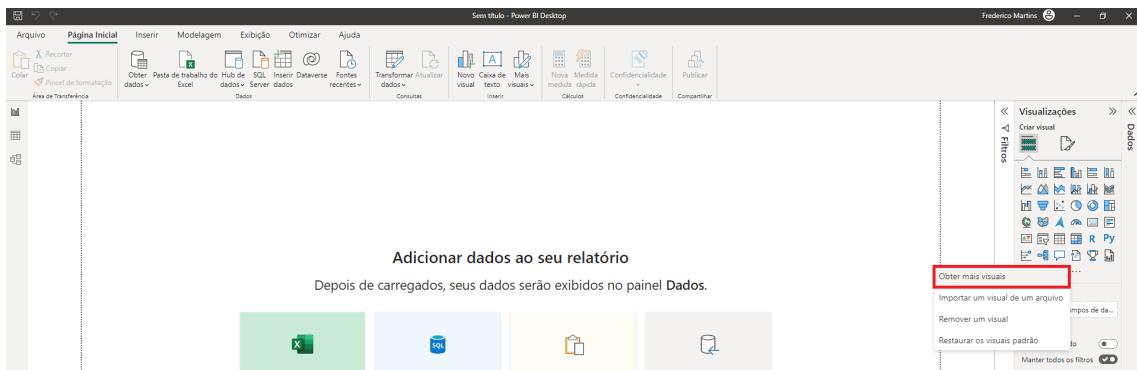
Pelos motivos citados anteriormente em conjunto com a facilidade da conexão com o datalake, optamos em utilizar o Microsoft Power BI que é uma ferramenta de Business Intelligence da Microsoft que possibilita, o tratamento de dados para a obtenção de informações relevantes, suportando análises e tomadas de decisão.

É possível ainda, através deste software, o desenvolvimento e manutenção de sistemas de monitoramento de performance, através de indicadores de desempenho.

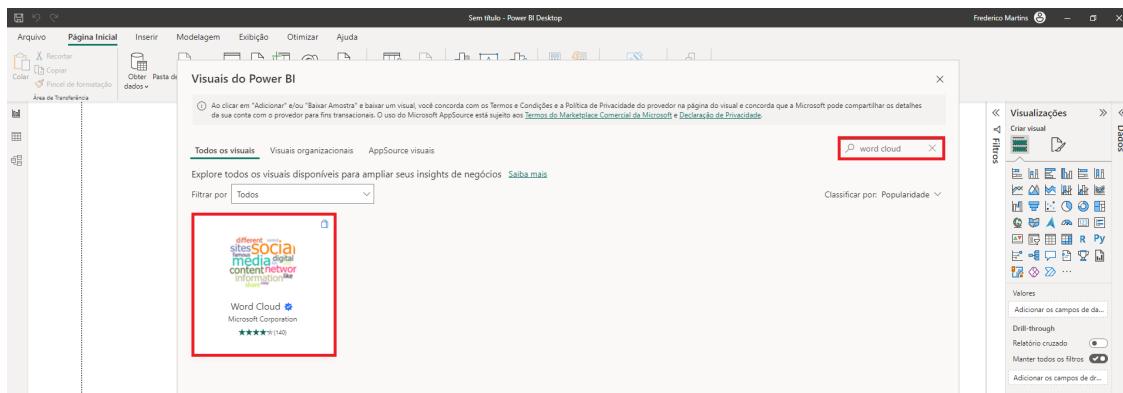
5.2 Instalação Gráfico WordCloud

Obtendo visual de Word Cloud, clique em obter mais visuais, três pontinhos:





Digite word cloud e selecione o visual Word Cloud:



Clique em Adicionar

Word Cloud
Microsoft Corporation
4.2 (140)
PBI Certified

Visão geral **Classificações + revisões**

Create a fun visual from frequent text in your data. Word Cloud is a visual representation of word frequency and value. Use it to get instant insight into the most important terms in your data. With the interactive experience of Word Cloud in Power BI, you no longer have to tediously dig through large volumes of text to find out which terms are prominent or prevalent. You can simply visualize them as Word Cloud and get the big picture instantly and user Power BI's interactivity to slice and dice further to uncover the themes behind the text context. This visual also puts you in control on the appearance of the word cloud, be it the size or usage of space and how to treat the data. You can choose to break the words in the text to look for the frequency word or keep word break off to project a measure as a value of the text. You can also enable stop words to remove the common terms from the word cloud to avoid the clutter. By enabling rotation and playing with the angles allowed, you can become very creative with this visual. Optionally you can also use a measure to provide weightage to the text. If none provided, it will simply use the frequency. Check out the formatting pane for more options. This is an open source visual. Get the code from GitHub: <https://github.com/Microsoft/PowerBI-visuals-wordcloud>

Funcionalidades do visual
Este visual é certificado pelo Power BI.
[Saiba mais sobre visuais certificados do Power BI.](#)

Adicionar
Baixar Amostra
[Instruções de amostra](#)

Preços
Gratuito

Produtos
Visuais do Power BI

Editor
Microsoft Corporation

Adquirir Usando
Conta corporativa ou de estudante

Versão
2.0.0

Atualizado
10/03/2022

Suporte
Suporte
Ajuda

Informações legais
Contrato de Licença
Política de Privacidade

Visual obtido, basta clicar no ícone do visual WordCloud 2.0.0 para utilizar.

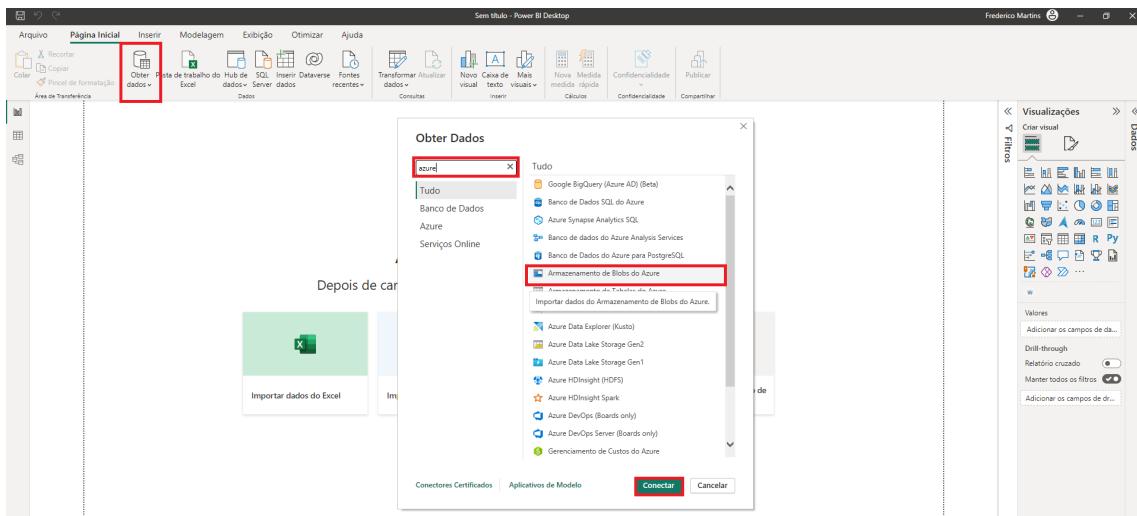
Arquivo **Página Inicial** **Inserir** **Modelagem** **Exibição** **Otimizar** **Ajuda**

Inserir **Visualizações**

WordCloud 2.0

5.3 Conexão Datalake

Clique em obter dados, no campo de busca digite azure e selecione Armazenamento de Blobs Azure:



Abra o portal Azure e clique no DataLake criado no grupo de recursos:

Nome	Tipo	Última visualização
stdatalakepuc	Conta de armazenamento	24 horas atrás

Procure por Pontos de Extremidade (End Point) e clique:

Página inicial >

stdatalakepuc Conta de armazenamento

Pesquisar

Inventário de blobs

Site estático

Gerenciamento do ciclo de vida

Configurações

Configuração

Compartilhamento de recursos (CORS)

SFTP

Recomendações do assistente

Pontos de extremidade

Bloqueios

Carregar Abrir no Gerenciador

Fundamentos

Grupo de recursos (mover) : Projeto_Big_D

Localização : East US

Assinatura (mover) : Azure for Stu

ID da Assinatura : 6410eba2-f01

Estado do disco : Disponível

Rótulos (editar) : Clique aqui para adi

Propriedades Monitoramento F

Data Lake Storage Namespace hierárquico

Navegue pelos pontos de extremidade até serviço Blob e copie o link:

Página inicial > stdatalakepuc

stdatalakepuc | Pontos de extremidade Conta de armazenamento

Pesquisar Atualizar

Estado de provisionamento : Êxito

Criado : 25/03/2023, 6:00:17 PM

ID do recurso da conta de armazenamento : /subscriptions/6410eba2-f013-49e5-8d2b-25aed89d7c90/resourceGroups/Projeto_Big_Data/providers/Microsoft.Storage/storageAccounts/stdatalakepuc

Serviço Blob

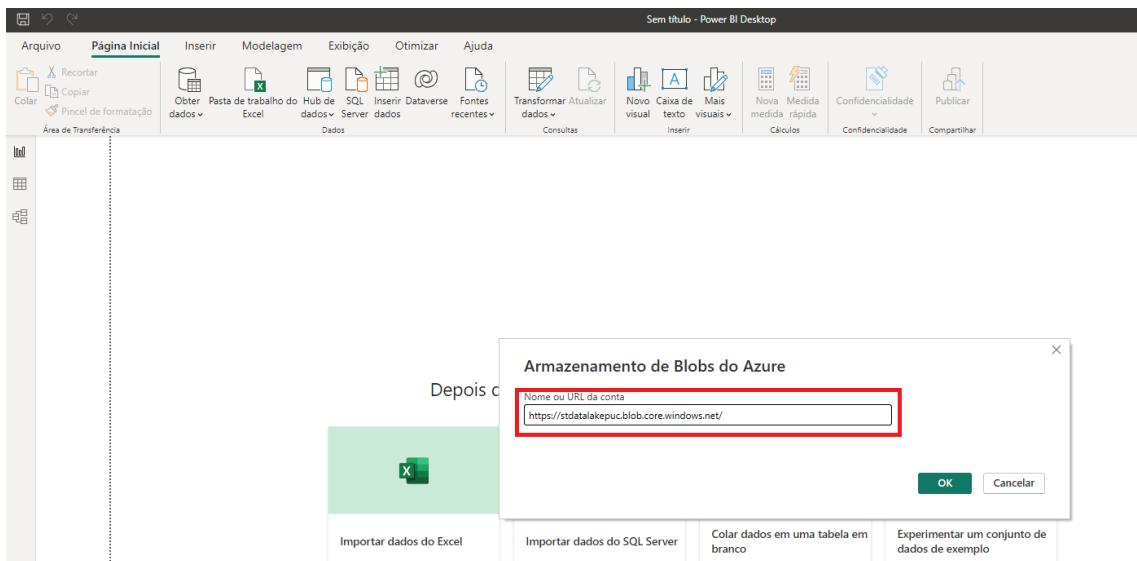
ID do Recurso : /subscriptions/6410eba2-f013-49e5-8d2b-25aed89d7c90/resourceGroups/Projeto_Big_Data/providers/Microsoft.Storage/storageAccounts/stdatalakepuc Copiar para área de | Copiado

Serviço Blob : https://stdatalakepuc.blob.core.windows.net/

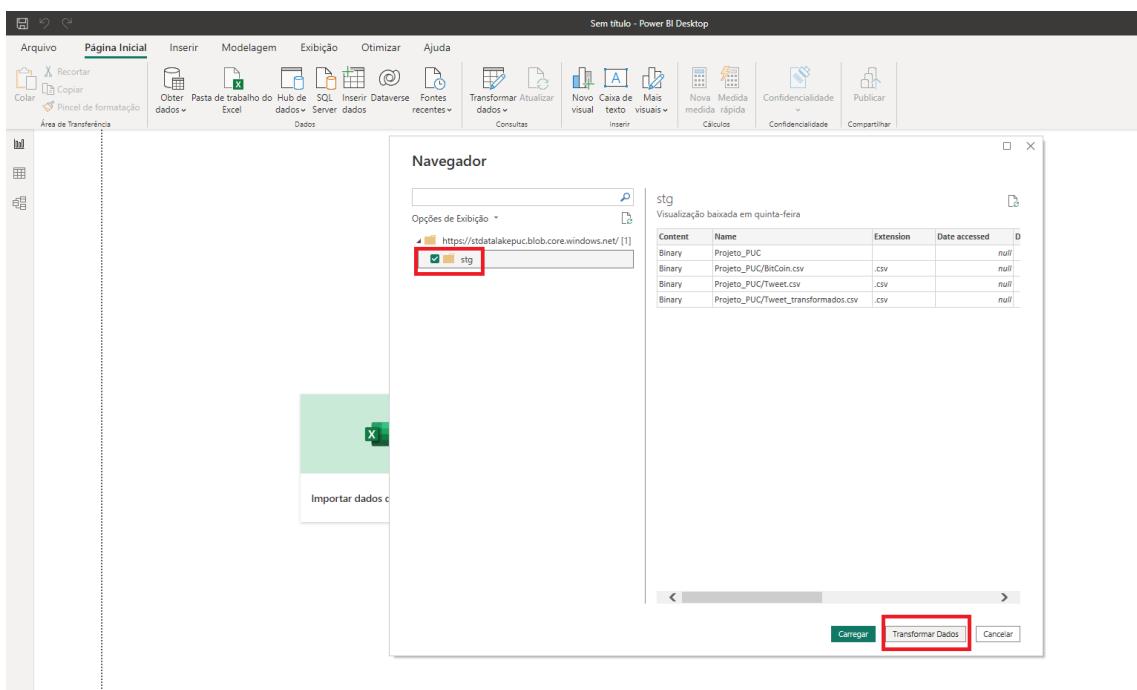
Serviço de arquivo

ID do Recurso : /subscriptions/6410eba2-f013-49e5-8d2b-25aed89d7c90/resourceGroups/Projeto_Big_Data/providers/Microsoft.Storage/storageAccounts/stdatalakepuc/fileSer...

Voltando no PowerBI cole o link copiando em Armazenamento de Blobs do Azure e clique em ok:



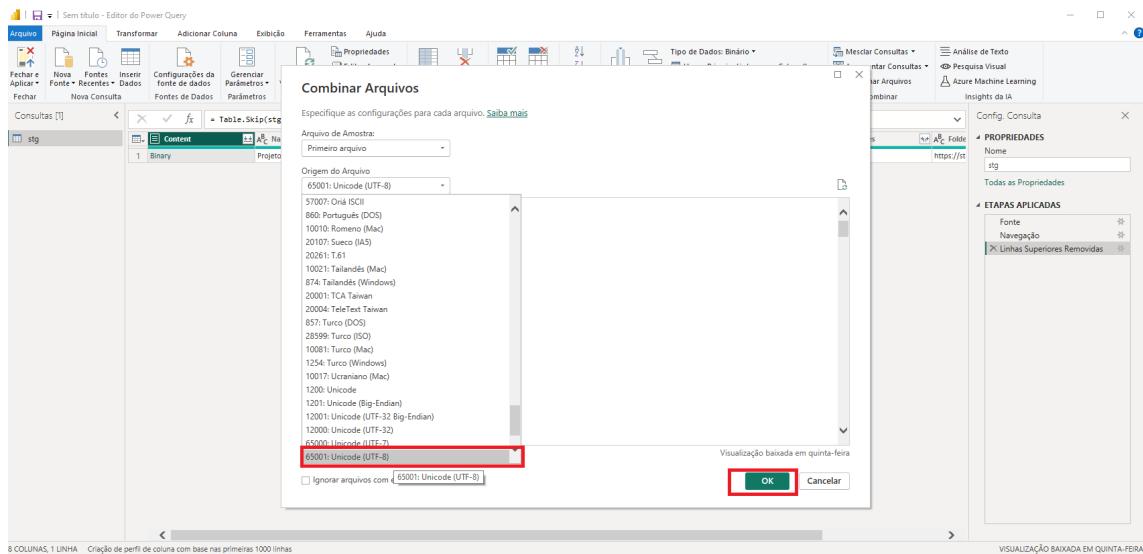
Selecione a pasta STG e clique em Transformar Dados:



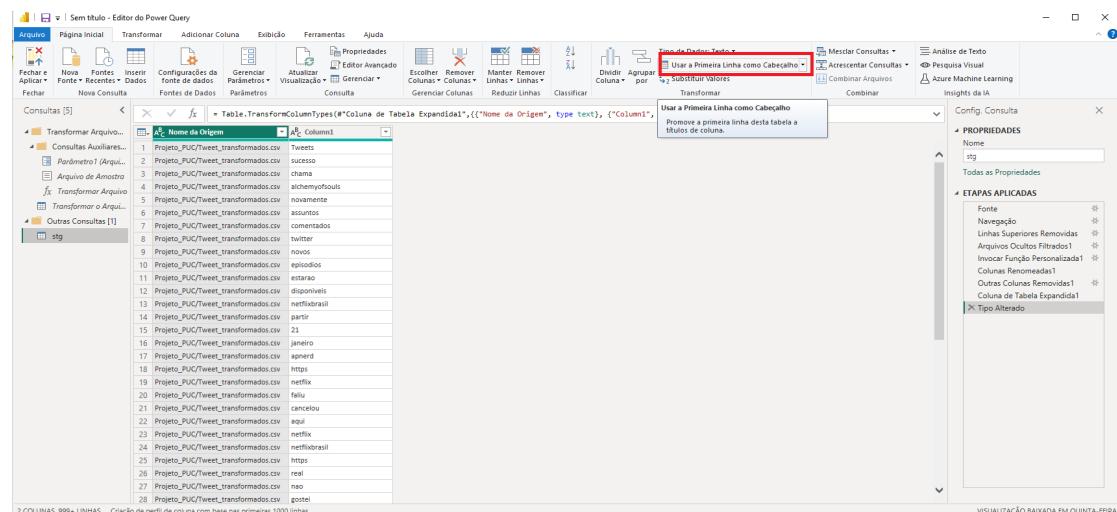
Cada linha é um arquivo que está no Data Lake, elimine as linhas não
são referentes a tabela desejada:

Na linha definida clique em Combinar Arquivos:

Selecione a Origem do Arquivo como UTF-8 e clique em ok:



Clique em usar a primeira linha de cabeçalho para adicionar o nome nas colunas:



Clique em Fechar e Aplicar:

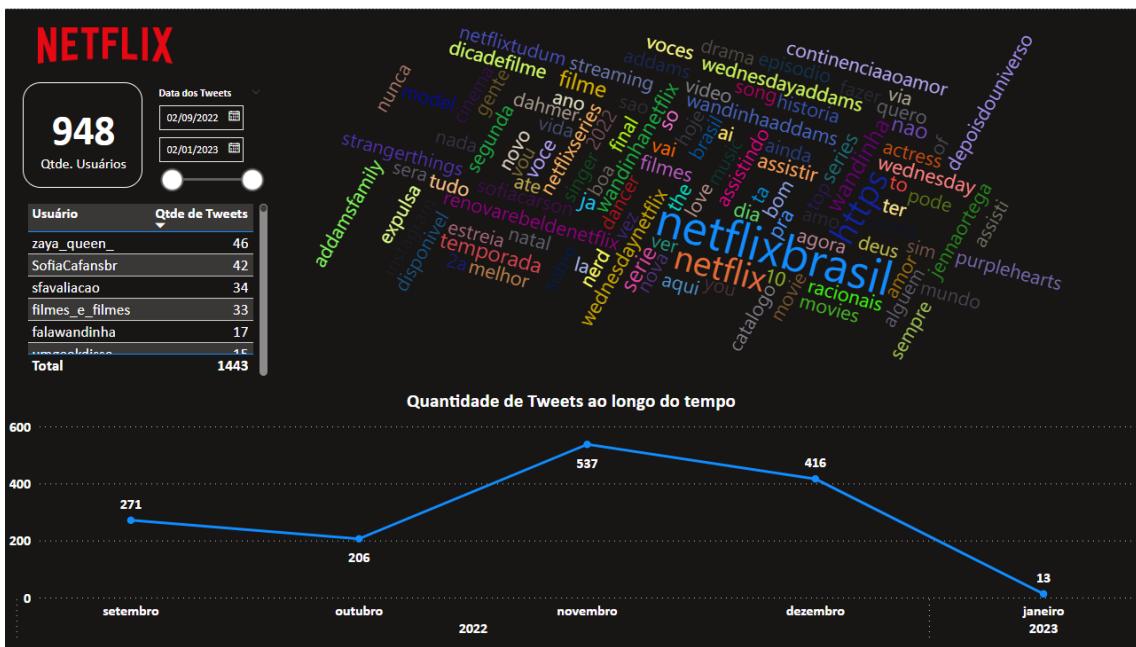
The screenshot shows the Power Query Editor interface. At the top, there's a ribbon with tabs like Arquivo, Página Inicial, Transformar, Exibição, Ferramentas, and Ajuda. Below the ribbon, there's a toolbar with various icons for file operations, data transformations, and visualizations. The main area displays a list of transformation steps applied to a table named 'Tweets'. The steps include reading from a CSV file, changing column types, and applying various transformations. A sidebar on the right shows the 'Configurações de Consulta' (Query Settings) and the 'ETAPAS APLICADAS' (Applied Steps) pane, which lists the transformations performed. The status bar at the bottom indicates '2 COLUMNAS, 999+ LINHAS'.

Pronto, conexão estabelecida com o Data Lake e tabela vinculada:

The screenshot shows the Power BI Desktop interface with the 'stg' table loaded. The 'Ferramentas da tabela' (Table Tools) tab is selected. The table view shows a large dataset with many rows and columns. The 'Dados' (Data) pane on the right shows the schema of the 'stg' table, which includes a single column named 'Tweets'. The status bar at the bottom indicates 'Tabela: stg (20.843 linhas)'.

5.4 Resultado da utilização do Power BI

Abaixo o views do insight obtido com a coleta de dados ao longo do projeto.



6. Conclusão

Este projeto teve como objetivo principal coletar feedbacks relacionados a repressão ao compartilhamento de senhas de usuários que utilizam o serviço de Streaming da Netflix.

A análise ocorreu no período 3 meses entre 02/09/2022 à 02/01/2023, com 948 usuários e totalizando 1443 tweets coletados.

Através da análise dos insights, não tivemos tweets no período estipulado que fosse relacionado ao principal assunto deste projeto. Os resultados obtidos foram de séries, filmes e estreias relevantes que estavam acontecendo naquele período.

Porém a medida entrou em vigor em 23/05/2023 no Brasil e em outros locais, passando a ser válido futuramente obter novos insights e resultados, após a implantação desta nova regra para identificar se houve alteração no comportamento e nos feedbacks de usuários.

Referências bibliográficas

COOBAN, Anna. Netflix restringe compartilhamento de senhas em mais quatro países. **CNN**, Londres, 2023. Disponível em:

<https://www.cnnbrasil.com.br/business/netflix-restringe-compartilhamento-de-senhhas-em-mais-quatro-paises/>. Acesso em: 11 mar. 2023.

Quezada, Vicent. **Big data, a aposta do mercado para o crescimento no Brasil.** 2021. Disponível em: <https://www.computerweekly.com/br/reportagen/Big-data-a-aposta-do-mercado-para-o-crescimento-no-Brasil>. Acesso em: 26 fev. 2023.