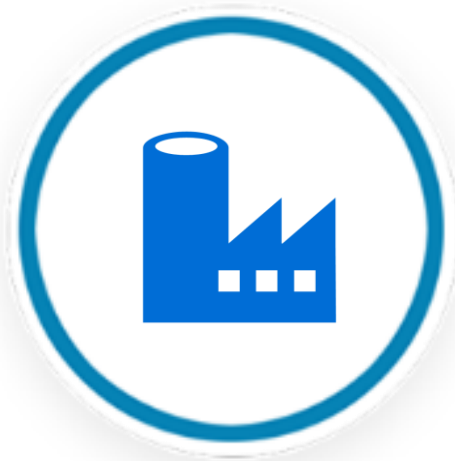

Tutorial

Criação de um Data Factory usando a interface do usuário



Azure Data Factory

1. Links Úteis	3
2. Descrição	3
2.1 Pré-requisitos	3
3. Criando os arquivos e pastas de entrada	3
4. Criar um data factory	5
5. Criar um serviço vinculado	7
6. Criar conjuntos de dados	9
7. Criar um pipeline	11
8. Depurar o pipeline	13
9. Disparar o pipeline manualmente	14
10. Monitorar o Pipeline	14
11. Disparar o pipeline em um cronograma	15
12. Considerações finais	17

1. Links Úteis

- [Documentação](#)

2. Descrição

É um serviço de integração de dados com baseado em nuvem que permite que você crie fluxos de trabalho orientados a dados na nuvem para orquestrar e automatizar a movimentação de dados e a transformação de dados. Usando o Azure Data Factory, é possível criar e agendar fluxos de trabalho orientados a dados (chamados de pipelines) que podem ingerir dados de diferentes repositórios de dados. Ele pode processar e transformar dados usando serviços de computação como o Azure HDInsight Hadoop, Spark, Azure Data Lake Analytics e Azure Machine Learning.

2.1 Pré-requisitos

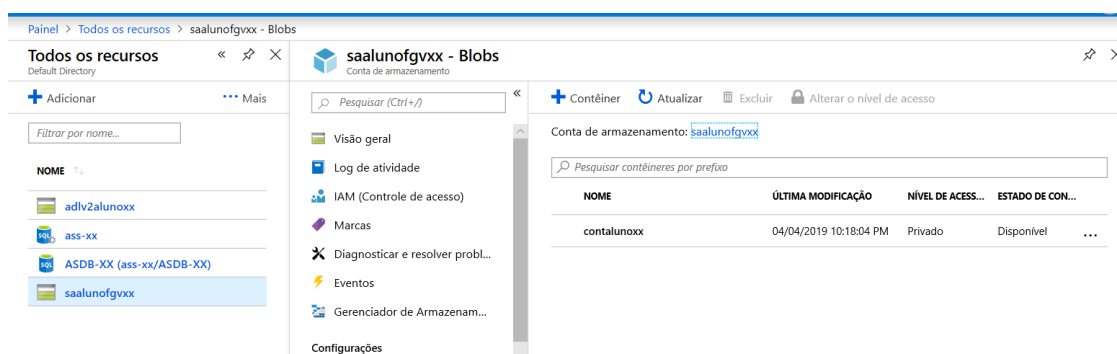
Para concluir este tutorial, verifique se está instalado:

- ✓ Possui uma conta no portal do Azure (<http://portal.azure.com>)

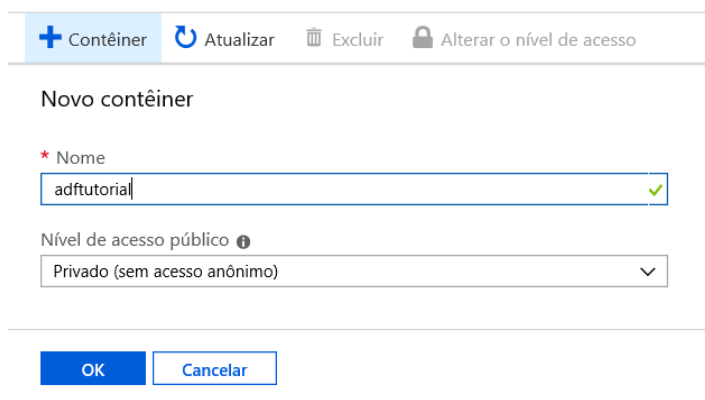
3. Criando os arquivos e pastas de entrada

Nesta seção, você cria um contêiner de blobs chamado **adftutorial** no armazenamento de Blobs do Azure. Você cria uma pasta chamada **entrada** no contêiner e, em seguida, carrega um arquivo de exemplo na pasta de entrada.

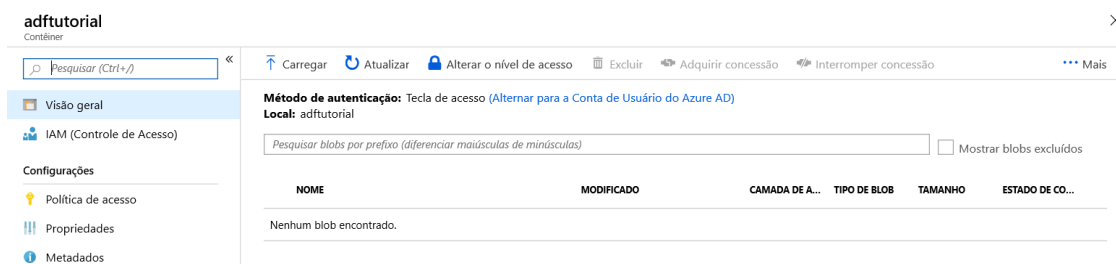
- Na janela **Conta de armazenamento**, alterne para **Visão geral** e depois selecione **Blobs**.
- Na página **Serviço Blob**, selecione + **Contêiner** na barra de ferramentas.



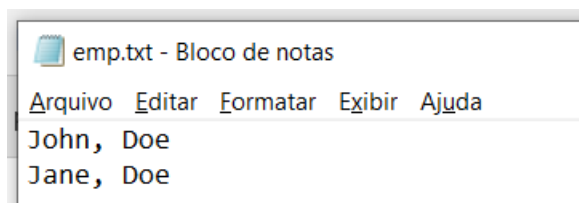
- Na caixa de diálogo **Novo contêiner**, insira **adftutorial** como o nome e selecione **OK**.



- Selecione **adftutorial** na lista de contêineres.
- Na página **Contêiner**, selecione **Carregar** na barra de ferramentas.



- f) Na página **Carregar blob**, selecione **Avançado**.
- g) Inicie o **Bloco de notas** e crie um arquivo chamado **emp.txt** com o seguinte conteúdo. Salve-o na pasta desejada.




- h) No Portal do Azure, na página **Carregar blob**, procure e selecione o arquivo **emp.txt** para a caixa **Arquivos**.
- i) Insira **entrada** como um valor da caixa **Carregar para a pasta**.
- j) Confirme que a pasta é **entrada** e o arquivo é **emp.txt** e selecione **Carregar**.

Carregar blob

adftutorial/

Arquivos ⓘ



☐ Substituir caso os arquivos já existam

^ Avançado

Tipo de Autenticação ⓘ

Conta de usuário do Azure AD

Chave de conta

Tipo de blob ⓘ

Blob de blocos

▼

☒ Carregar arquivos .vhd como blobs de páginas (recomendado)

Tamanho do bloco ⓘ

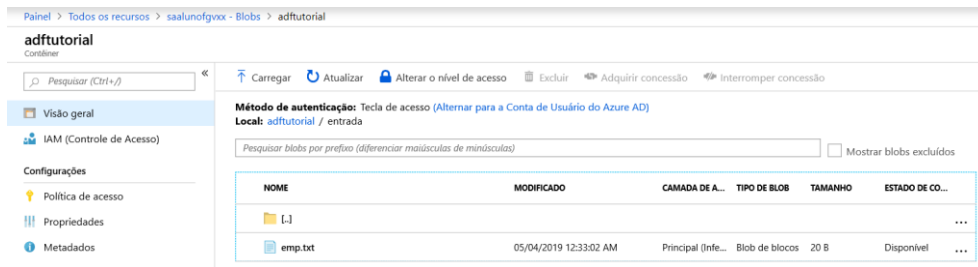
4 MB

▼

Carregar na pasta

Carregar

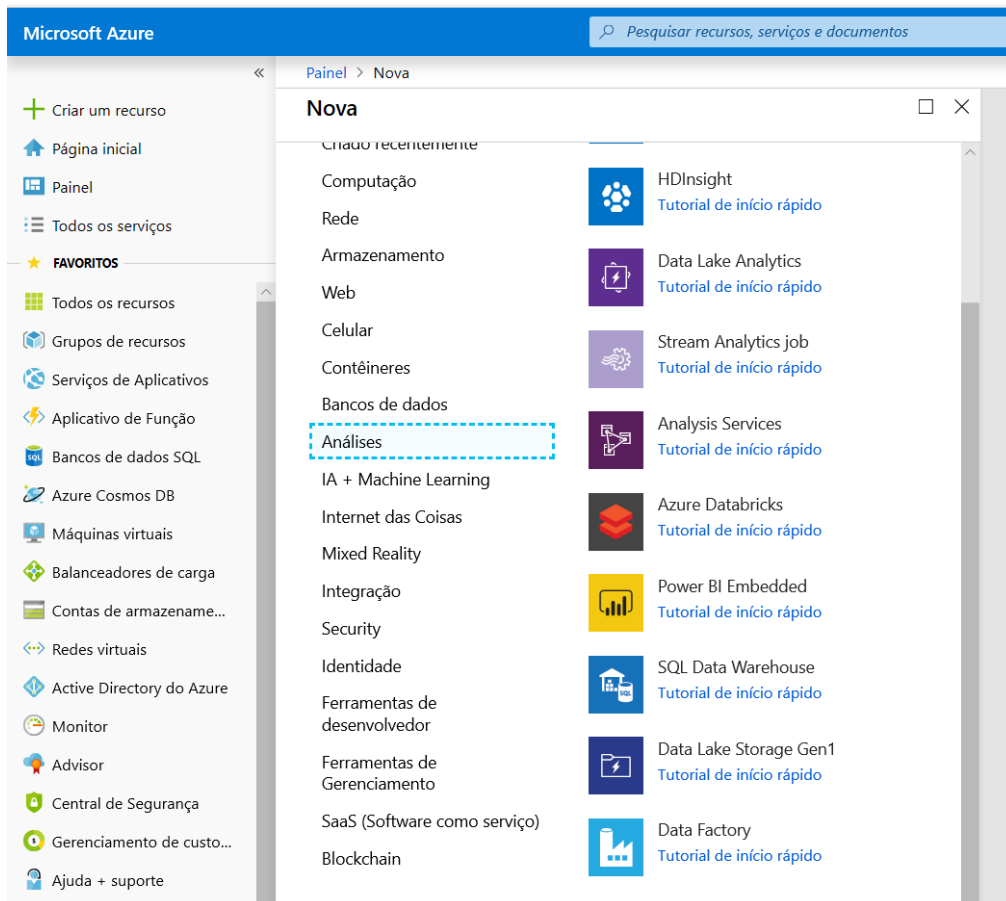
- k) O arquivo **emp.txt** e o status do carregamento devem estar na lista.



- l) Feche a página **Carregar blob** clicando no **X** no canto superior.
- m) Mantenha a página **Contêiner** aberta. Você a usa para verificar a saída no final do guia de início rápido.

4. Criar um data factory

- a) Iniciar o navegador da Web **Microsoft Edge** ou **Google Chrome**. Atualmente, a interface do usuário do Data Factory tem suporte apenas nos navegadores da Web Microsoft Edge e Google Chrome.
- b) Vá para o [Portal do Azure](#).
- c) Selecione **Criar recurso** no menu à esquerda, depois **Análise** e, por fim, **Data Factory**.



- d) Na página **Novo data factory**, insira **ADF** no campo **Nome**.
- e) O nome do Azure Data Factory deve ser *globalmente exclusivo*. Se você visualizar o seguinte erro, altere o nome de data factory (por exemplo, **AlunoFGVXXADF**) e tente criar novamente. Para ver as regras de nomenclatura para artefatos do Data Factory consulte o artigo [Data Factory - regras de nomenclatura](#).
- f) Para **Assinatura**, selecione a assinatura do Azure na qual você deseja criar o data factory.

- g) Para o **Grupo de Recursos**, use uma das seguintes etapas:
 - a. Selecione **Usar existente** e selecione um grupo de recursos existente na lista.
- h) Para **Versão**, selecione **V2**.
- i) Em **Local**, selecione uma localização para o data factory.
- j) A lista mostra somente os locais aos quais o Data Factory dá suporte e em que os metadados do Azure Data Factory serão armazenados. Observe que os armazenamentos de dados (como o Armazenamento do Azure e o Banco de Dados SQL do Azure) e serviços de computação (como o Azure HDInsight) usados pelo Data Factory podem ser executados em outras regiões.
- k) Selecione **Criar**.

Painel > Nova > New data factory

New data factory

* Name ⓘ
AlunoFGVXXADF ✓

* Assinatura
Avaliação Gratuita ▼

* Resource Group ⓘ
☐ Criar novo ☒ Usar existente
 RG-XX ▼

Version ⓘ
V2 ▼

* Localização ⓘ
Leste dos EUA 2 ▼

Integrate with GIT source control to do collaboration, source control, change tracking, change difference, continuous integration and deployment etc

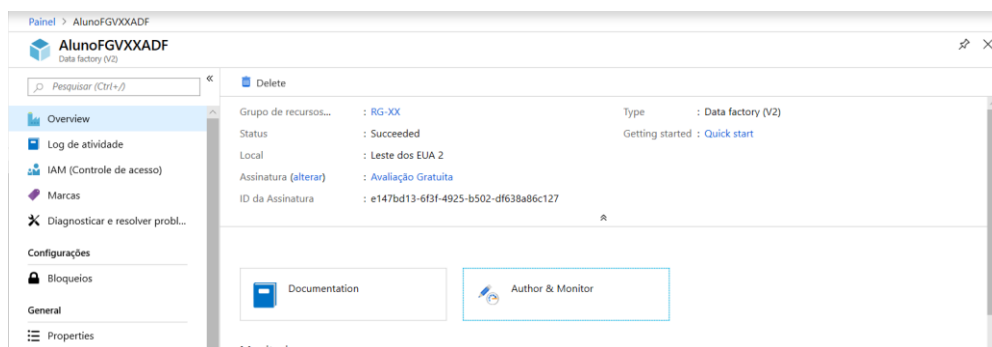
☐ Enable GIT ⓘ

* GIT URL ⓘ
https://github.com/michaelbond

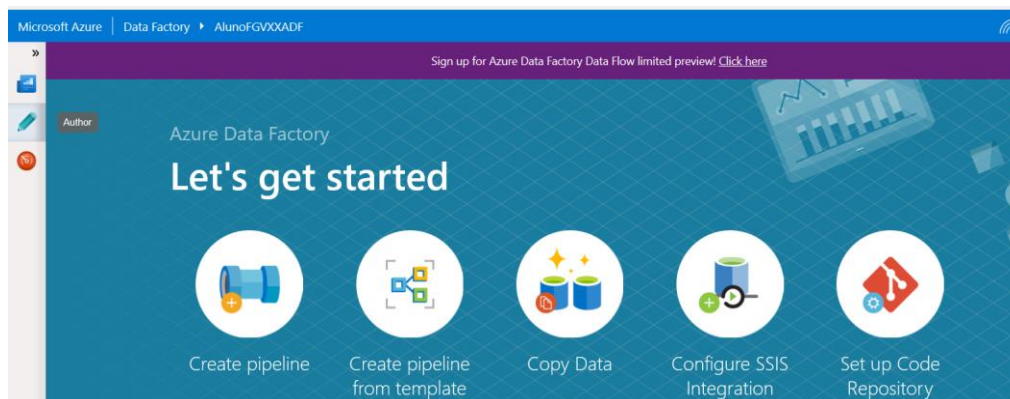
* Repo name ⓘ

Criar Opções de automação

- l) Após a criação, a página do **Data Factory** será exibida. Clique no bloco **Criar e Monitorar** para iniciar o aplicativo de interface do usuário (IU) do Azure Data Factory em uma guia separada.



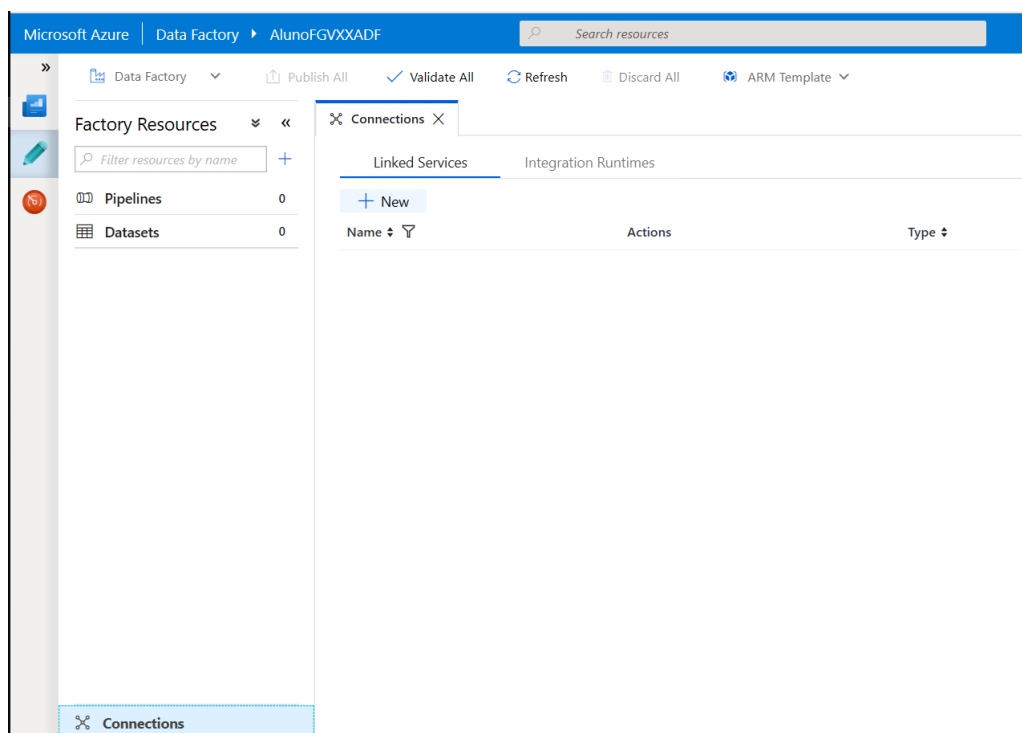
m) Na página **Introdução**, acesse a guia **Autor** no painel esquerdo.



5. Criar um serviço vinculado

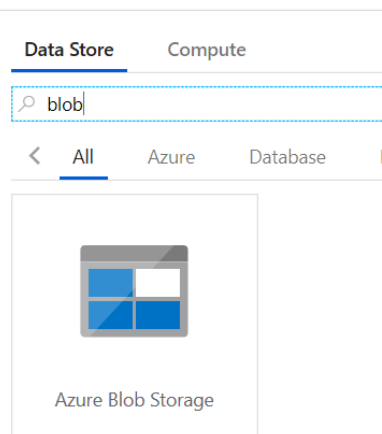
Nesta etapa, você criará um serviço vinculado para vincular sua Conta de Armazenamento do Azure ao Data Factory. O serviço vinculado tem as informações de conexão que o serviço do Data Factory usa no tempo de execução para se conectar a ele.

a) Clique em **Conexões** e, em seguida, selecione o botão **Novo** na barra de ferramentas.



- b) Na página **Novo Serviço Vinculado**, selecione **Armazenamento de Blobs do Azure** e selecione **Continuar**.

New Linked Service



- c) Conclua as seguintes etapas:
- Para o campo **Nome**, insira **LS_ saalunofgvxx**.
 - Insira o nome da sua Conta de Armazenamento do Azure em **Nome da conta de armazenamento**.
 - Selecione **Testar conectividade** para confirmar se o serviço do Data Factory pode se conectar à conta de armazenamento.
 - Para salvar o serviço vinculado, selecione **Concluir**.

←

New Linked Service (Azure Blob Storage)

×

Name *

LS_saalunofgvxx

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication method

Account key

Connection String

Azure Key Vault

Account selection method

☒ From Azure subscription
 ☐ Enter manually

Azure subscription

Avaliação Gratuita (e147bd13-6f3f-4925-b502-df638a86c127)

Storage account name *

saalunofgvxx

Additional connection properties

+ New

Annotations

+ New

✓ Connection successful

Cancel

Test connection

Finish

6. Criar conjuntos de dados

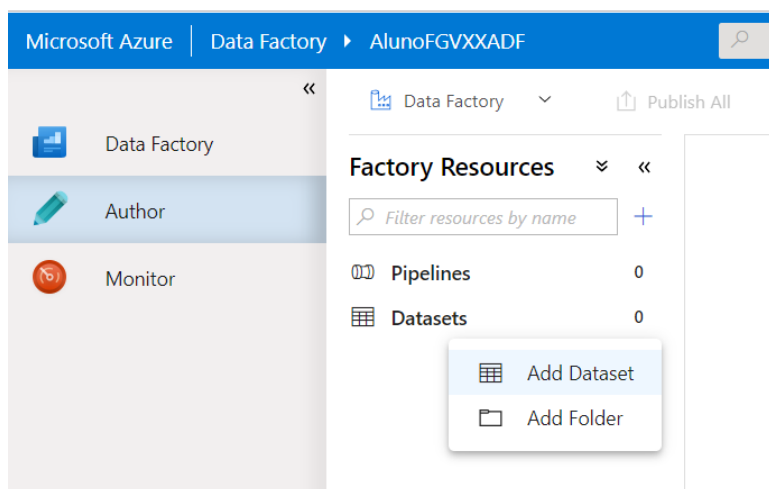
Neste procedimento, você criará dois conjuntos de dados: **InputDataset** e **OutputDataset**. Esses conjuntos de dados são do tipo **AzureBlob**. Eles se referem ao Serviço vinculado do Armazenamento do Azure que você criou na seção anterior.

O conjunto de dados de entrada representa os dados de origem na pasta de entrada. Na definição de conjunto de dados de entrada, especifique o contêiner de blob (**adftutorial**), a pasta (**entrada**) e o arquivo (**emp.txt**) que contém os dados de origem.

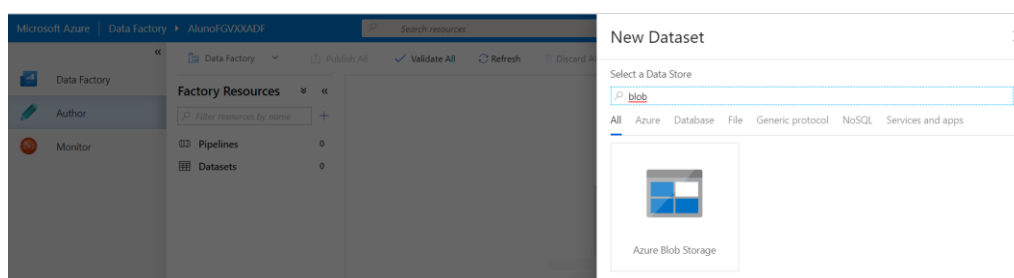
Esse conjunto de dados de saída representa os dados que são copiados para o destino. Na definição de conjunto de dados de saída, especifique o contêiner de blob (**adftutorial**), a pasta (**saída**) e o arquivo para o qual os dados são copiados. Cada execução de um pipeline tem uma ID exclusiva associada a ele. Você pode acessar essa ID, usando a variável do sistema **RunId**. O nome do arquivo de saída é avaliado dinamicamente com base na ID de execução do pipeline.

Nas configurações de serviço vinculado, você especificou a conta de armazenamento do Azure que contém os dados de origem. Nas configurações do conjunto de dados de origem, especifique onde exatamente os dados de origem residem (contêiner de blob, pasta e arquivo). Nas configurações de conjunto de dados do coletor, especifique para onde os dados são copiados (contêiner de blob, pasta e arquivo).

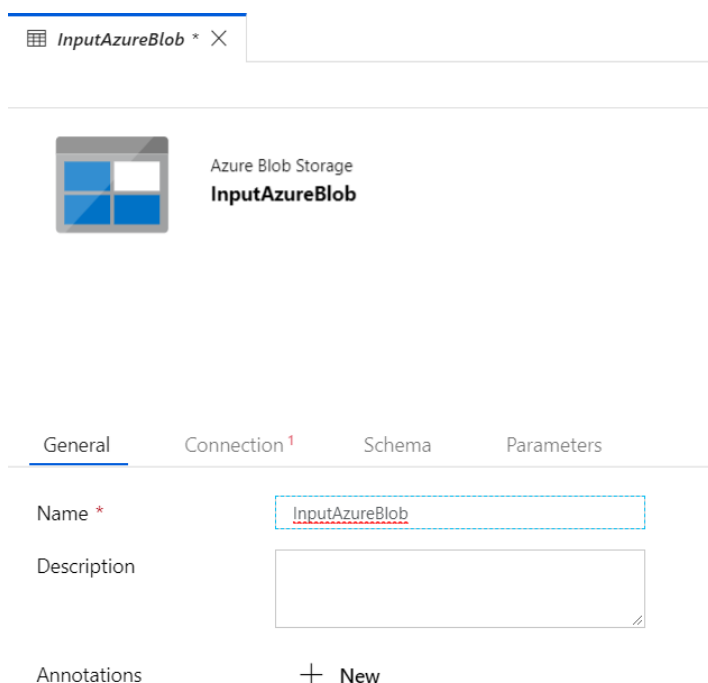
- Selecione o botão + (mais) adição e, em seguida, selecione **Conjunto de Dados**.



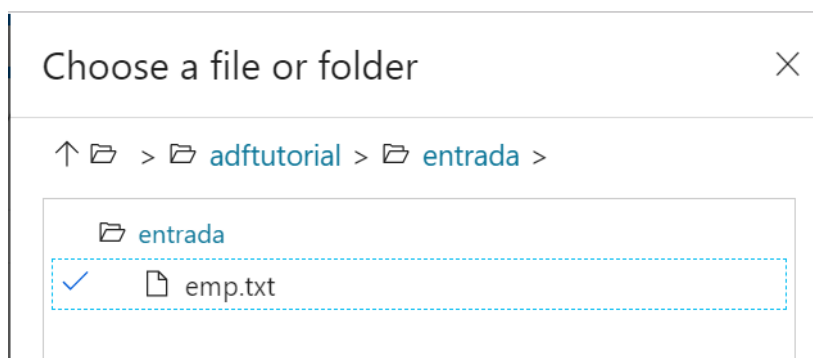
- b) Na página **Novo Conjunto de Dados**, selecione **Armazenamento de Blobs do Azure** e selecione **Concluir**.



- c) Na janela **Geral** do conjunto de dados, insira **InputDataset** para o **Nome**.



- d) Alterne para a guia **Conexão** e siga estas etapas:
- Selecione **AzureStorageLinkedService** para **Serviço vinculado**.
 - Para **Caminho do arquivo**, selecione o botão **Procurar**.
 - Na janela **Escolher um arquivo ou uma pasta**, navegue até a pasta **entrada** no contêiner **adftutorial**, selecione o arquivo **emp.txt** e clique em **Concluir**.



- d. (opcional) Selecione **Visualizar dados** para visualizar os dados no arquivo emp.txt.

Data Preview

Linked Service: LS_saalunofgvxx

Object: adftutorial/entrada/emp.txt

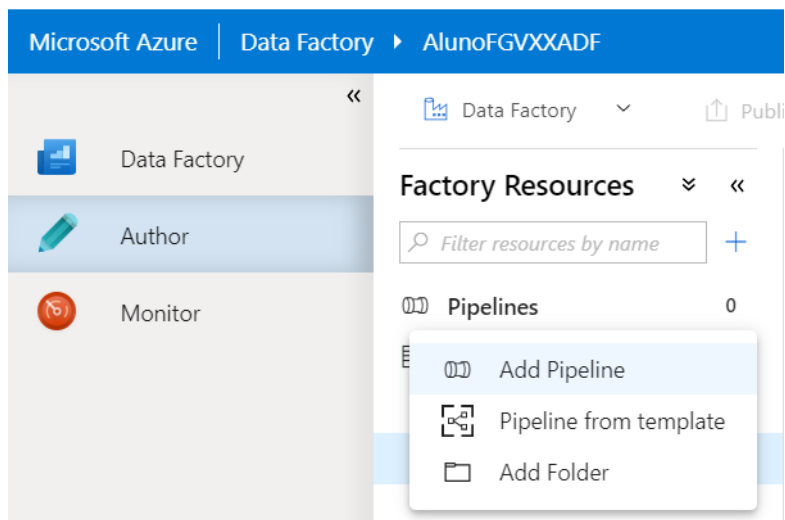
Prop_0	Prop_1
John	Doe
Jane	Doe

- e) Repita as etapas para criar o conjunto de dados de saída:
- Selecione o botão + (mais) adição e, em seguida, selecione **Conjunto de Dados**.
 - Na página **Novo Conjunto de Dados**, selecione **Armazenamento de Blobs do Azure** e selecione **Concluir**.
 - Na tabela **Geral**, especifique **OutputDataset** para o nome.
 - Na guia **Conexão**, selecione **AzureStorageLinkedService** como serviço vinculado e, no campo de diretório, insira **adftutorial/output** para a pasta. Se a pasta de **saída** não existir, a atividade de cópia a cria em tempo de execução.

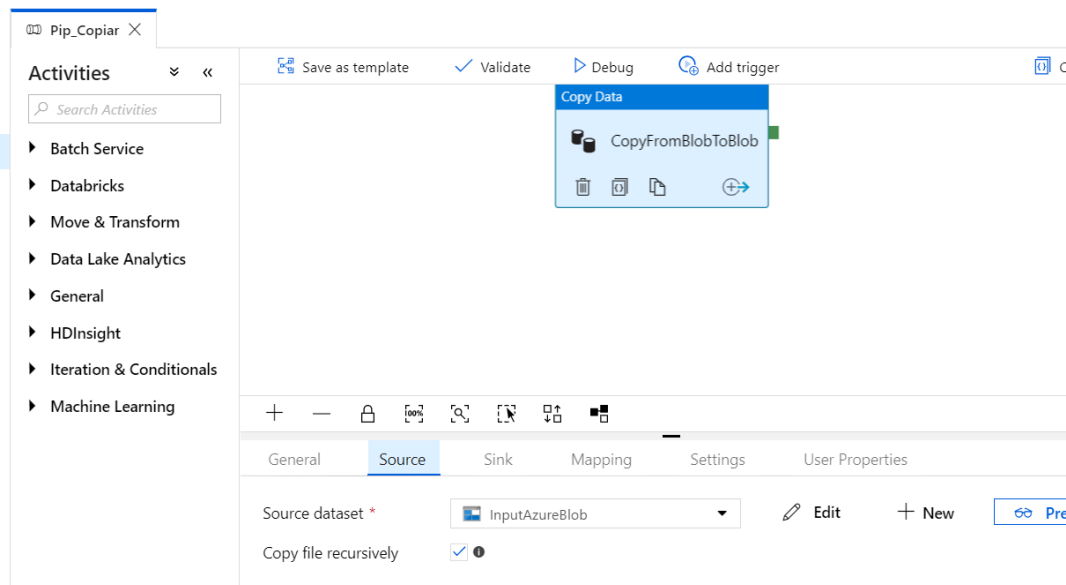
7. Criar um pipeline

Neste procedimento, você criará e validará um pipeline com uma atividade Copiar que usa o conjunto de dados de entrada e saída. A Atividade de cópia copia os dados do arquivo especificado por você nas configurações do conjunto de dados de entrada para o arquivo especificado por você nas configurações do conjunto de dados de saída. Se o conjunto de dados de entrada especifica apenas uma pasta (não o nome de arquivo), a Atividade de cópia copia todos os arquivos da pasta de origem para o destino.

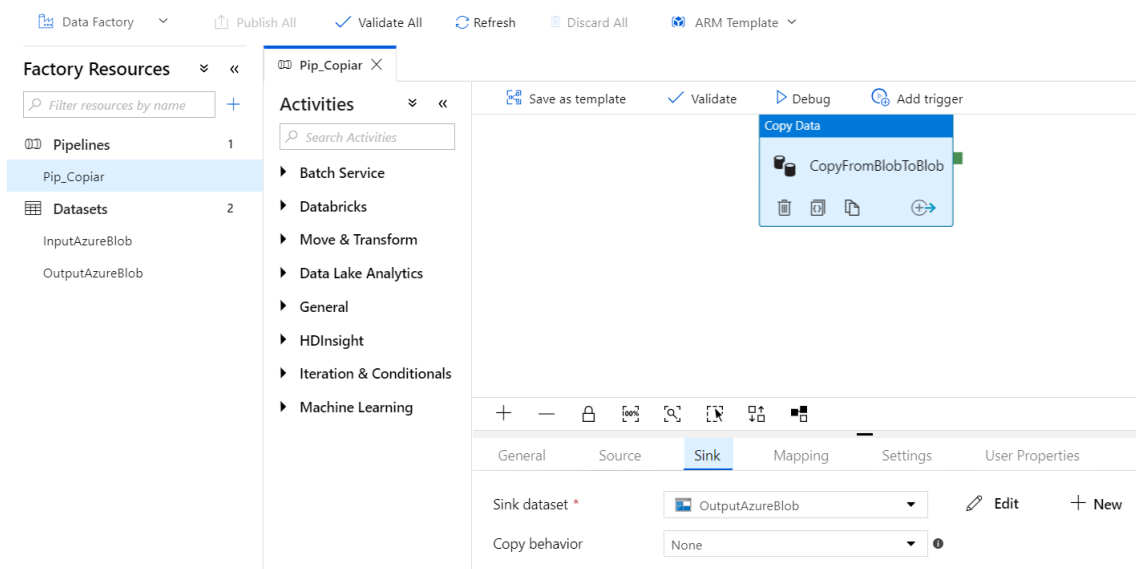
- Selecione o botão + (mais) adição e, em seguida, selecione **Pipeline**.



- b) Na guia **Geral**, especifique **CopyPipeline** para o **nome**.
- c) Na caixa de ferramentas **Atividades**, expanda **Mover e transformar**. Arraste e solte a atividade de **Cópia** da caixa de ferramentas **Atividades** para a superfície do designer do pipeline. Você também pode pesquisar atividades na caixa de ferramentas **Atividades**. Especifique **CopyFromBlobToBlob** para o **Nome**.
- d) Alterne para a guia **Fonte** nas configurações da atividade de cópia e selecione **InputDataset** para o **Conjunto de dados de origem**.



- e) Alterne para a guia **Coletor** nas configurações da atividade de cópia e selecione **OutputDataset** para o **Conjunto de dados do coletor**.



- f) Clique em **Validar** na barra de ferramentas do pipeline sobre a tela para validar as configurações de pipeline. Confirme se esse pipeline foi validado com êxito. Para fechar a saída de validação, selecione o botão >> (seta para a direita).

Pipeline Validation Output

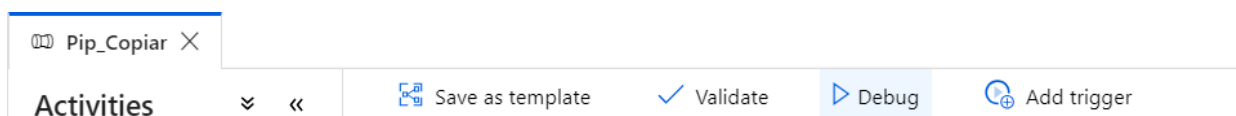


Your Pipeline has been validated.
No errors were found.

8. Depurar o pipeline

Nesta etapa, você depura o pipeline antes de implantá-lo no Data Factory.

- a) Na barra de ferramentas do pipeline acima da tela, clique em **Depurar** para disparar uma execução de teste.



- b) Confirme se você vê o status da execução do pipeline na guia **Saída** das configurações do pipeline na parte inferior.

General

Parameters

Variables

Output

Pipeline Run ID: 152e3648-fd66-4d9a-b051-611903e03020 [🔗] [🔄] ⓘ

NAME	TYPE	RUN START	DURATION	STATUS	ACTIONS	RUNID
CopyFromBlo...	Copy	04/05/2019 8:13 PM	00:00:05	✔ Succeeded	📄 📁 🔍	3d959b

- c) Confirme que você vê um arquivo de saída na pasta **saída** do contêiner **adftutorial**. Se a pasta de saída não existir, o serviço do Data Factory a cria automaticamente.

adftutorial

↶

↷

UploadDownload

↶

+

OpenNew Folder

🔗

Copy URL

📄

☑

Select All

📄

📄

CopyPaste

🏷

Rename

✕

↶

DeleteUndelete

📷

📷

Create SnapshotManage Snapshots

↶

↷

⌵

⬆

Active blobs (default)

adftutorial > saida

Search

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Status	Remaining Days
emp.txt	Hot (inferred)		05/04/2019 20:13:23	Block Blob	application/octet-stream	20 B	Active	

9. Disparar o pipeline manualmente

Nesta procedimento, você implanta entidades (serviços vinculados, conjuntos de dados, pipelines) ao Azure Data Factory. Depois, dispare manualmente a execução do pipeline.

- Antes de disparar um pipeline, você deve publicar as entidades no Data Factory. Para publicar, selecione **Publicar Tudo** na parte superior.
- Para disparar o pipeline manualmente, selecione **Gatilho** na barra de ferramentas do pipeline e selecione **Disparar Agora**.

adftutorial												
Upload	Download	Open	New Folder	Copy URL	Select All	Copy	Paste	Rename	Delete	Undelete	Create Snapshot	Manage Snapshots
Active blobs (default) adftutorial > saída										Search by prefix (case-sensitive)		
Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Status	Remaining Days	Deleted Time	Lease S		
emp.txt	Hot (inferred)		05/04/2019 20:16:51	Block Blob	application/octet-stream	20 B	Active					

10. Monitorar o Pipeline

- Altere para a guia **Monitorar** à esquerda. Use o botão **Atualizar** para atualizar a lista.

Microsoft Azure

Data Factory

AlunoFGVXXADF

Data Factory

Author

Monitor

Dashboards

Pipeline Runs

Trigger Runs

Integration Runtimes

Alerts & Metrics

Run

Cancel options

Refresh

Last 24 Hours

04/04/2019 7:44 PM - 04/05/2019 7:44 PM

Time Zone

(UTC-03:00) Brasilia

View All Rerun History

All

Succeeded

In Progress

Queued

Failed

Cancelled

Pipeline Name

Actions

Run Start

Duration

Triggered By

Status

Parameters

Pip_Copiar

04/05/2019, 8:16:44 PM

00:00:07

Manual trigger

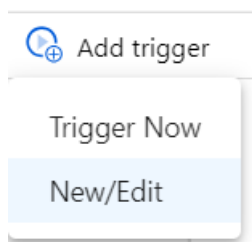
Succeeded

- b) Selecione **Exibir Execuções de Atividade** em **Ações**. Você vê o status de execução da atividade de cópia nesta página.
- c) Para exibir detalhes sobre a operação de cópia, selecione **Detalhes** (imagem de óculos) na coluna **Ações**. Para obter detalhes sobre as propriedades, confira [Visão geral da atividade de cópia](#).
- d) Confirme que você vê um arquivo novo na pasta de **saída**.
- e) Você pode alternar de volta para o modo de exibição **Execuções do pipeline** no modo de exibição **Execuções de Atividade** selecionando **Pipelines**.

11. Disparar o pipeline em um cronograma

Este procedimento é opcional neste tutorial. Você pode criar um *agendador de gatilho* para agendar a execução periódica do pipeline (por hora, diariamente, e assim por diante). Nesta procedimento, você cria um gatilho para ser executado a cada minuto até a data e hora de término especificadas.

- a) Alterne para a guia **Autor**.
- b) Vá até o pipeline, selecione **Gatilho** na barra de ferramentas do pipeline e depois selecione **Novo/Editar**.



- c) Na página **Adicionar gatilhos**, selecione **Escolher gatilho** e, em seguida, selecione **Novo**.
- d) Na página **Novo gatilho**, no campo **Final**, selecione **Na Data**, especifique como hora de término alguns minutos após a hora atual e selecione **Aplicar**.

Um custo associado a cada execução de pipeline, então, especifique o a hora de término como apenas alguns minutos após a hora de início. Verifique se está como o mesmo dia. No entanto, verifique se há tempo suficiente para a execução do pipeline entre a hora da publicação e a hora de término. O gatilho só entra em vigor depois de você publicar a solução no Data Factory, e não ao salvar o gatilho na interface do usuário.

- a) Na página **Novo gatilho**, selecione a caixa de seleção **Ativado** e, em seguida, selecione **Avançar**.
- b) Examine a mensagem de aviso e selecione **Concluir**.

← New Trigger
×

Name *

Description

Type *
☒ Schedule
☐ Tumbling Window
☐ Event

Start Date (UTC) *

Recurrence *
 Minute(s)

End *
☒ No End
☐ On Date

Annotations
+ New

☒ Activated

- c) Clique em **Publicar Tudo** para publicar as alterações no Data Factory.
- d) Alterne para a guia **Monitorar** à esquerda. Selecione **Atualizar** para atualizar a lista. Você verá que o pipeline é executado uma vez por minuto desde o momento da publicação até hora de término.
- e) Observe os valores na coluna **Disparado Por**. A execução do gatilho manual foi feita em uma etapa anterior (**Disparar agora**).
- f) Alterne para o modo de exibição **Execuções de gatilho**.
- g) Confirme que um arquivo de saída é criado para cada execução de pipeline até a data e hora de término especificadas na pasta **saída**.

12. Considerações finais

Neste tutorial, foi apresentado as seguintes tarefas:

- ✓ Criando os arquivos e pastas de entrada
- ✓ Criar um data factory
- ✓ Criar um serviço vinculado
- ✓ Criar conjuntos de dados
- ✓ Criar um pipeline
- ✓ Depurar o pipeline
- ✓ Disparar o pipeline manualmente
- ✓ Monitorar o Pipeline
- ✓ Disparar o pipeline em um cronograma

Esse material foi adaptado a partir do link original abaixo:

- ✓ <https://docs.microsoft.com/en-us/azure/sql-database/sql-database-design-first-database>

