PROJETO DATABRICKS SEGURO

MANUAL DE INSTALAÇÃO

Sumário

1 Introdução

O ambiente de execução desse projeto é provisionado no ambiente Azure Microsoft e requer a criação dos seguintes recursos:

- Uma conta de armazenamento para criação do ADSL (Azure Data Lake Storage);
- Um cofre de chaves (Azure Key Vault);
- Um workspace e um cluster do Azure Databricks.

Os recursos acima são provisionados de forma automatizada com o uso do ARM (Azure Resource Manager) e *scripts em Powershell,* e ficam organizados em um grupo de recursos com nome a ser definido em tempo de provisionamento.

2 Provisionamento da infraestrutura no Azure

O provisionamento dos recursos é feito através da execução de um script em Powershell. Os arquivos estão disponíveis no repositório do GitHub e devem ser baixados para uma estação de trabalho com sistema operacional Windows.

Atenção: esse documento está baseado na versão em português do Portal do Azure. Os *prints* de tela e menções aos componentes e recursos estão nesse idioma.

2.1 Pré-requisitos de execução

- É necessário ter um ambiente Azure acessível no momento da execução do script;
- O script deve ser executado em uma janela do *Powershell* com direitos administrativos;
- O script e o arquivo *template* devem estar na mesma pasta de armazenamento.

Essa infraestrutura pode ser provisionada tanto em ambientes de desenvolvimento e de produção da indústria, quanto em ambientes de pesquisa e para fins didáticos. É possível criar um ambiente Azure com créditos disponíveis para testes. Veja maiores informações em https://azure.microsoft.com/pt-br/free

Atenção: É desejável que você tenha alguma familiaridade com a utilização do Portal do Azure e execução de scripts Powershell, para acompanhar esse roteiro de instalação com mais facilidade.

2.2 Parametrização e execução do script de provisionamento dos recursos

O *script* e o *template* devem ser parametrizados com as seguintes informações:

Script:

- Nome do template que será implantado;
- Nome do grupo de recursos que será criado;
- Nome da conta de armazenamento criada;
- Nome do registro de aplicativo que será utilizado no acesso ao ADSL pelo Azure Databricks.

Atenção: As informações acima devem ser preenchidas nas variáveis no início do código do script. Isso pode ser feito em qualquer editor de texto (e.g. Notepad, Notepad ++, etc.)

Template:

- Nome do cofre de senhas;
- Nome da workspace do Azure Databricks;

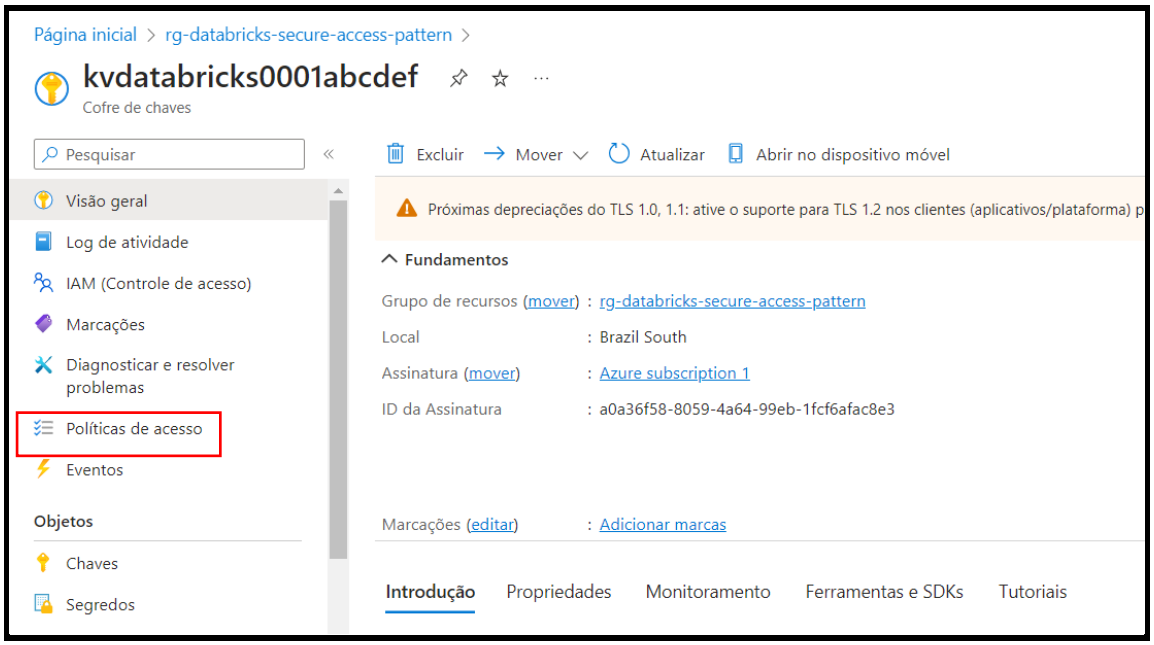
Para editar o template, use o arquivo template_base.json, substitua os textos <key vault name> e <databricks workspace name> pelos nomes do cofre de chaves e espaço de trabalho do Databricks, respectivamente. Salve o arquivo com o nome de template.json.

Atenção: os nomes da conta de armazenamento e do cofre de senhas devem ser exclusivos entre todos os existentes no Azure. Eles devem ter de 3 a 24 caracteres e podem conter somente letras minúsculas e números. Recomenda-se verificar os nomes válidos no Portal do Azure antes de preencher e executar o script de provisionamento.

2.3 Configurações de acesso e segredos no cofre de chaves

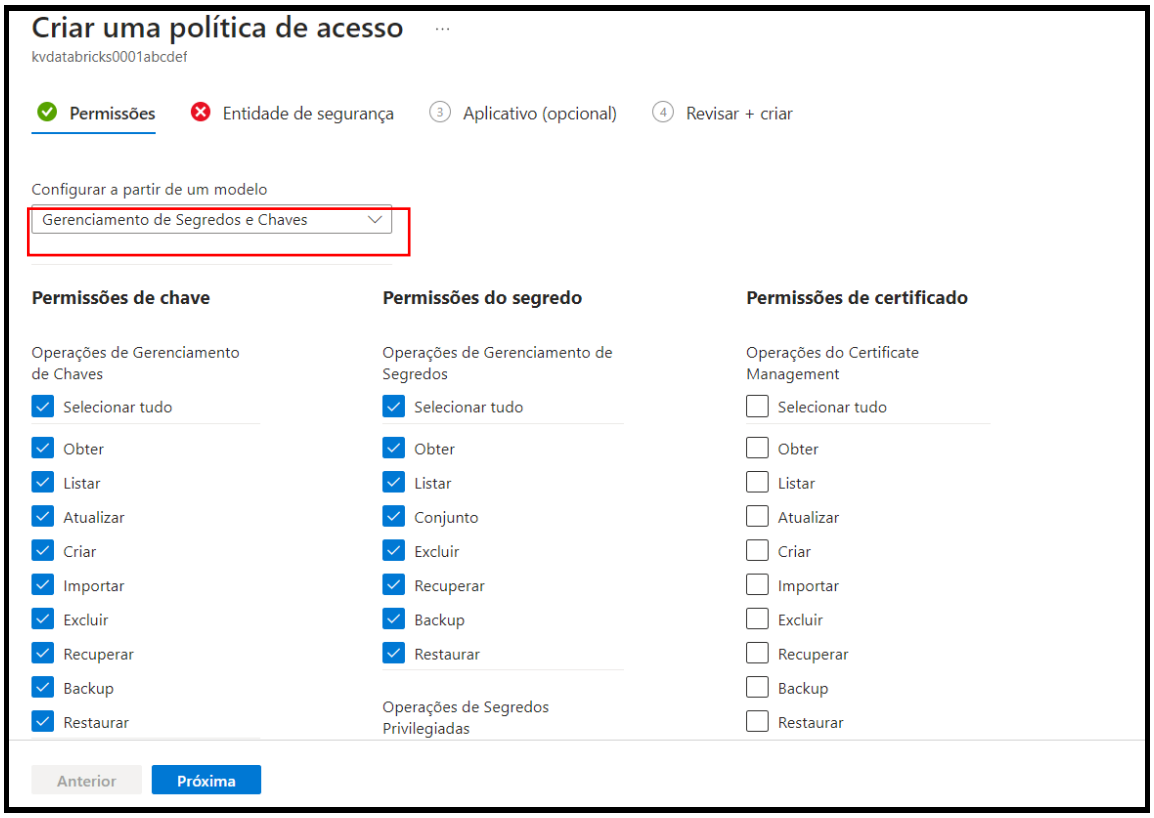
Por questões de segurança (exposição de informações sensíveis do locatário Azure), a configuração de acesso e a criação de segredos no cofre de chaves devem ser feitos diretamente no portal do Azure onde os recursos estão provisionados. Além disso, essas informações são dinâmicas e geradas em tempo de criação do respectivo recurso no locatário.

No Portal do Azure, localize o cofre de senhas criado após a implantação. Ele está dentro do grupo de recursos que você informou no processo de provisionamento. Nesse projeto, ele está com o nome de “kvdatabricks0001abcdef”.

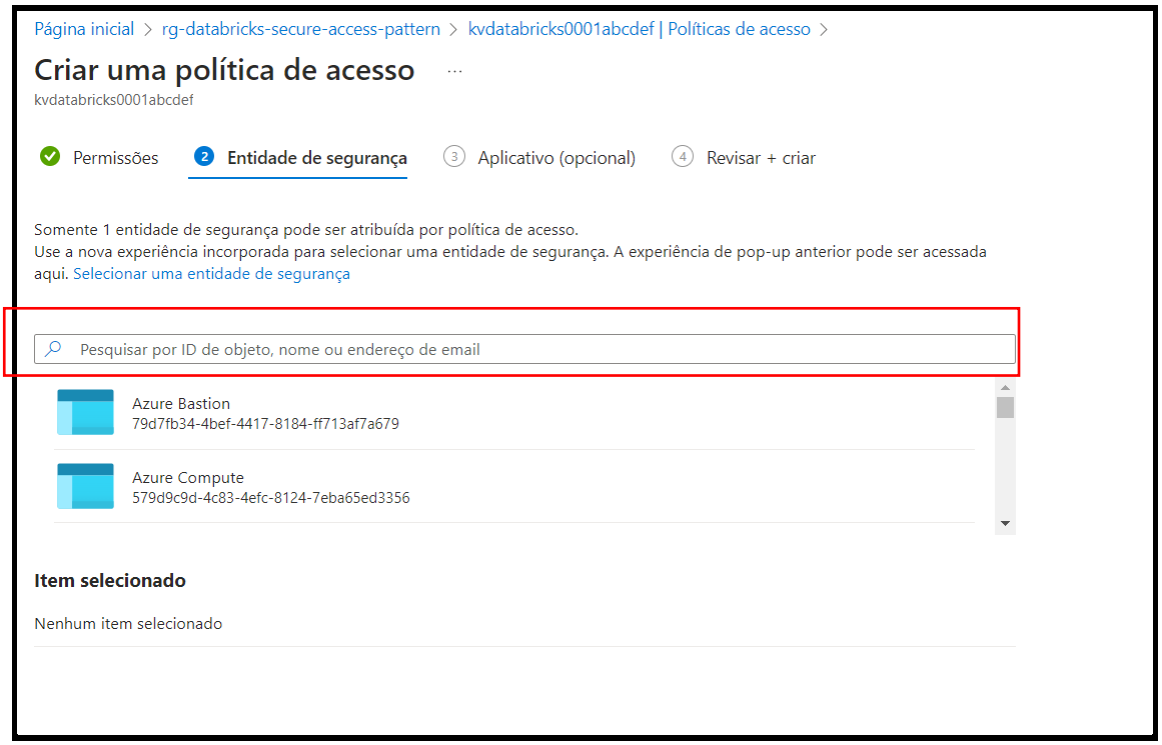


Clique na opção “Políticas de acesso”. Não haverá nenhuma política de acesso criada. Clique em “Adicionar Política” ou “+Criar”.

Em “Configurar a partir de um modelo” selecione “Gerenciamento de Segredos e Chaves” e clique em “Próxima”.



Pesquise pelo seu usuário na caixa indicada e selecione quando for localizado. Clique em "Próxima".

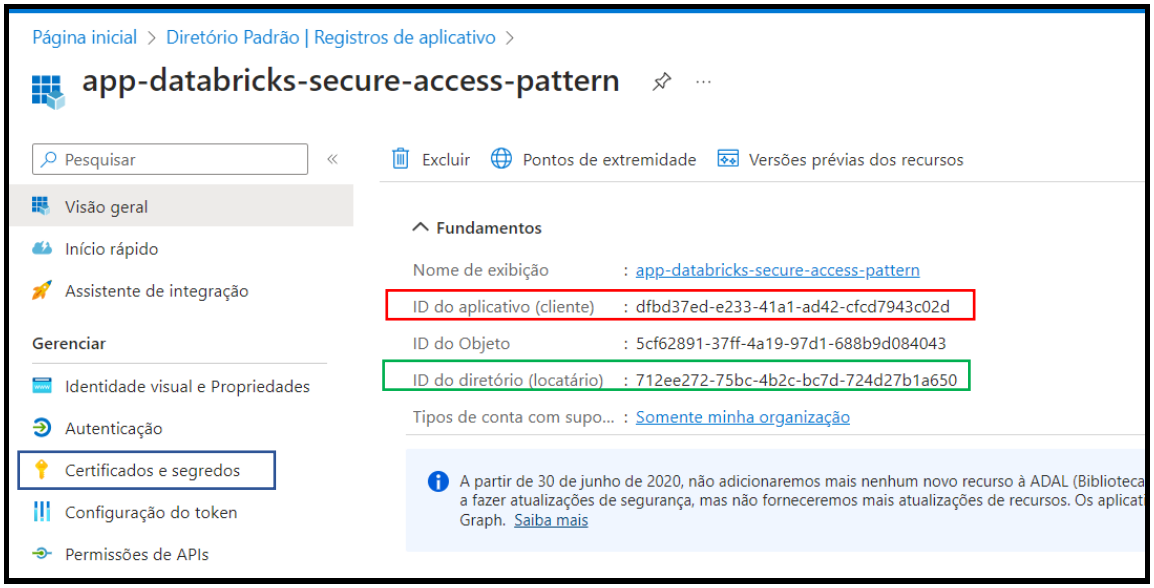


Na tela seguinte, continue sem selecionar nada, e por fim clique em “Criar”. Após essa configuração, será possível criar os segredos através do Portal.

Três segredos devem ser criados no cofre:

- client-id: deve possuir o ID de cliente do registro de aplicativo;
- tenant-id: deve possuir o ID o tenant do registro de aplicativo;
- client-secret: deve possuir o ID de cliente do registro de aplicativo.

Os valores de client_id e tenant-id são obtidos na visão geral do registro de aplicativo que foi provisionado no ambiente. Vá em Azure Active Diretory/Registros de Aplicativo, clique em “Todos os Aplicativos” e procure pelo nome do registro de aplicativo que você forneceu no script.

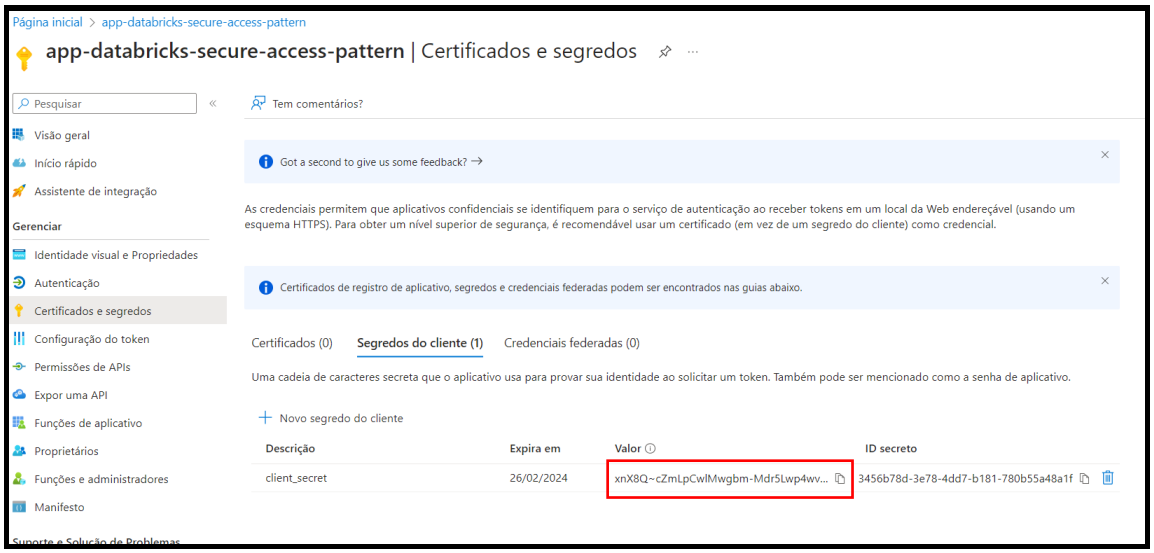


Nesse exemplo foi criado o registro de aplicativo “app-databricks-secure-access-pattern”. No retângulo vermelho está o valor do segredo do cliente-id e no retângulo verde está o valor do segredo do tenant-id. Esses valores serão utilizados para criar os segredos no cofre de chaves. Nessa mesma você irá gerar o client-secret. Clique em “Certificados e segredos” (retângulo azul).

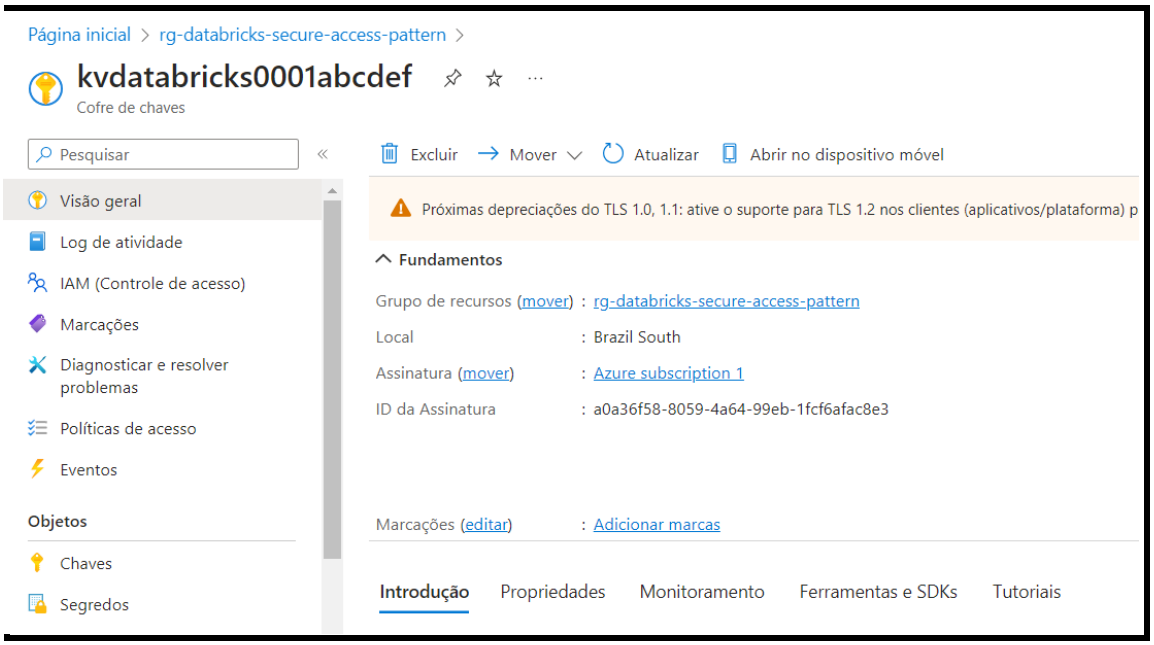
Na tela abaixo, clique em “Novo segredo do cliente”.

Preencha com o texto “client-secret” e clique em “Adicionar”. Ao voltar para a tela, observe que a informação no retângulo em vermelho é o valor do segredo do client-secret.

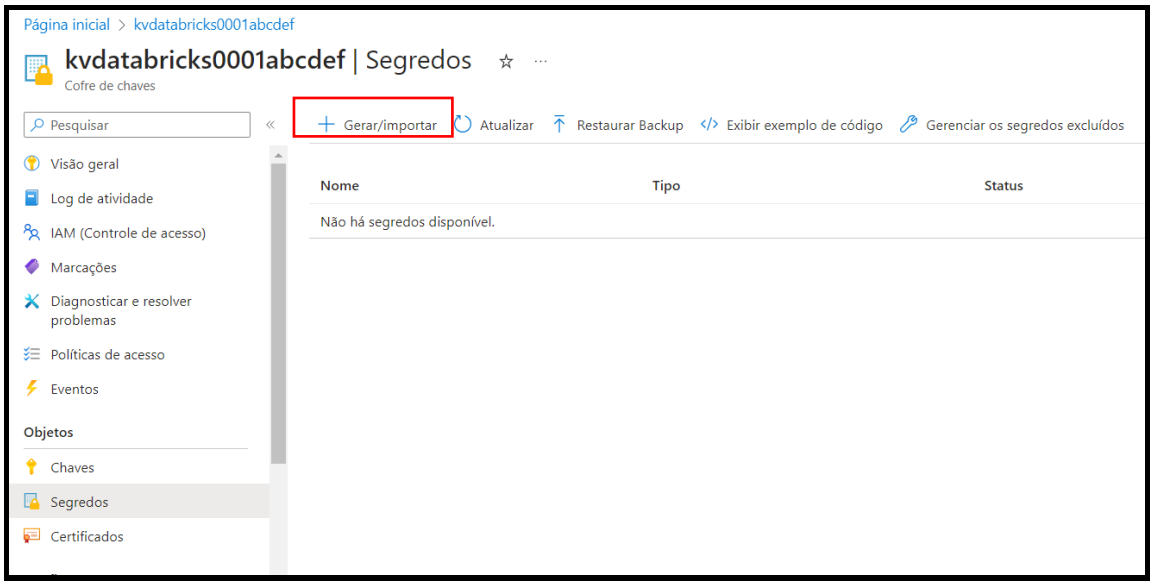
Atenção: faça essa configuração em uma aba específica do navegador, para que você possa recuperar essa informação na etapa de criação dos segredos no cofre. Caso você troque a tela, não terá mais a opção de copiar o valor, conforme aparece na tela abaixo.



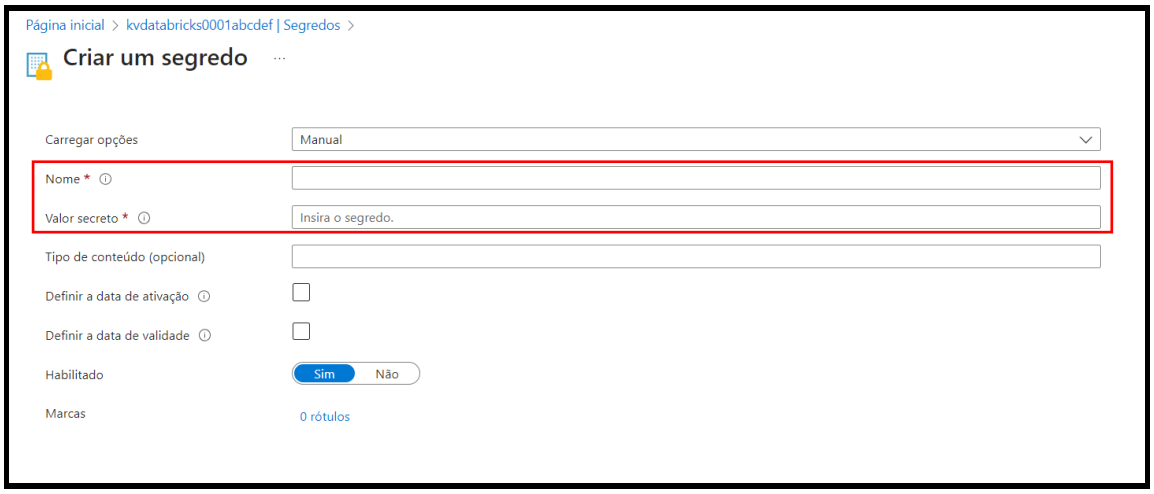
De posse dos segredos, é hora de configurá-los no cofre de senhas. Volte na tela de configuração do cofre de senhas e clique em “Segredos”.



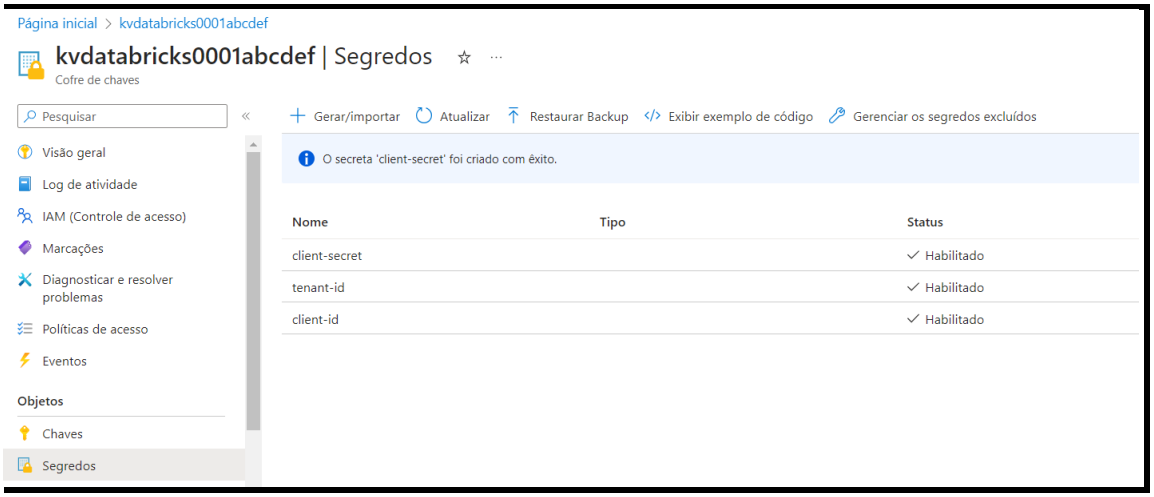
Clique em “Gerar/Importar”.



Na tela a seguir você irá cadastrar cada um dos três segredos que serão utilizados pelo Azure Databricks para acessar o Azure Data Lake Storage. No campo Nome você irá preencher com o nome do segredo (client-id, tenant-id e client-secret) e no campo Valor você irá colar a informação relacionada a esse campo.



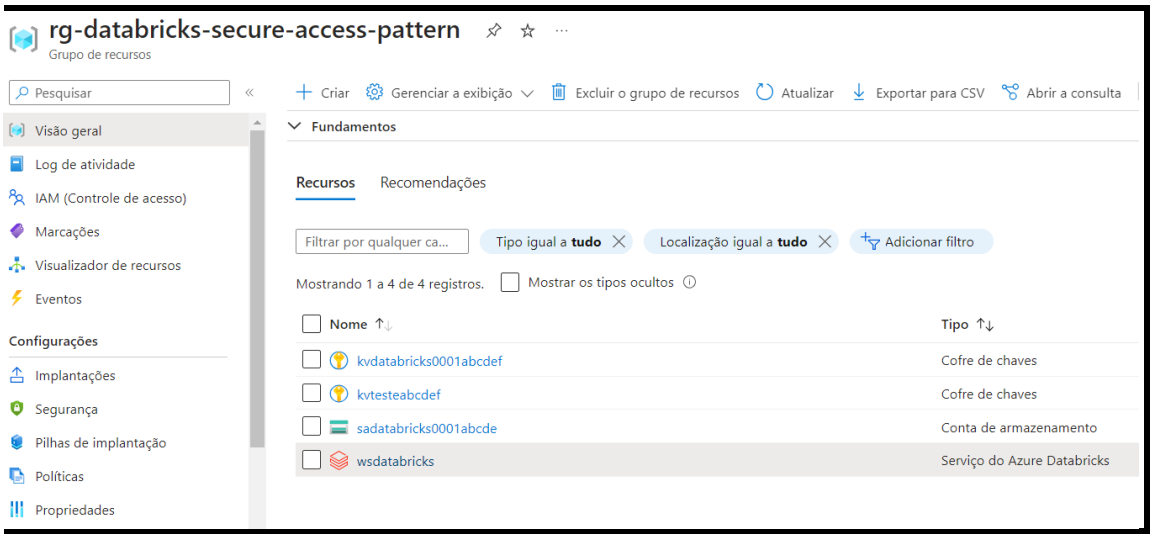
Ao final do processo, sua tele deve estar assim:



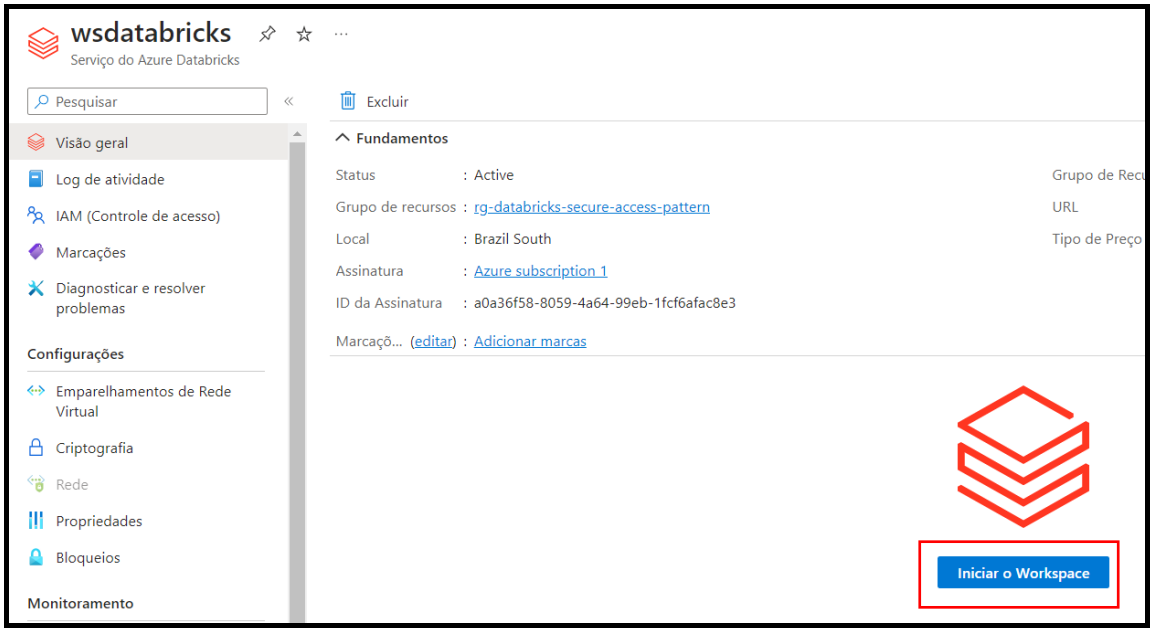
3 Configuração do Azure Databricks

3.1 Acesso ao Workspace

A workspace do Azure Databricks é um dos recursos criados de forma automática pelo script de provisionamento. O acesso a workspace pode ser feito através do Portal do Azure. Basta acessar o grupo de recursos recém-criado e selecionar o recurso do tipo “Serviço do Azure Databricks”



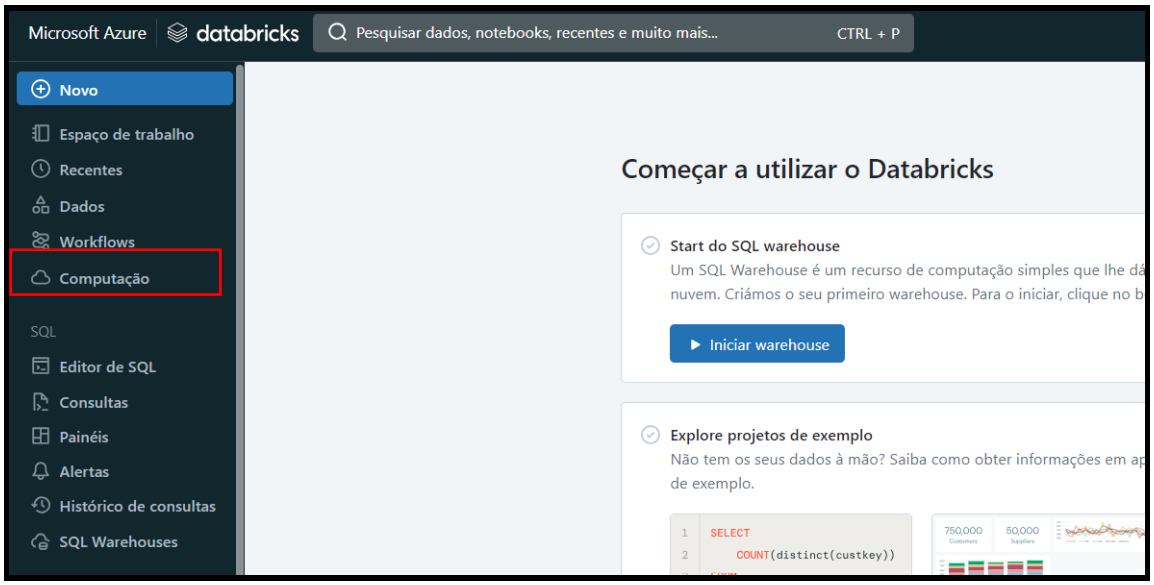
Clique nele e irá para a visão geral desse recurso. Clique em “Iniciar Workspace”.



Pronto! Agora você está na tela inicial da Workspace do Azure Databricks. É nesse ambiente que serão feitas as configurações finais para conexão com o Azure Data Lake Storage.

3.2 Criação do cluster

A primeira etapa do processo envolve a criação de um cluster. Na tela inicial do Azure Databricks escolha a opção “Computação”

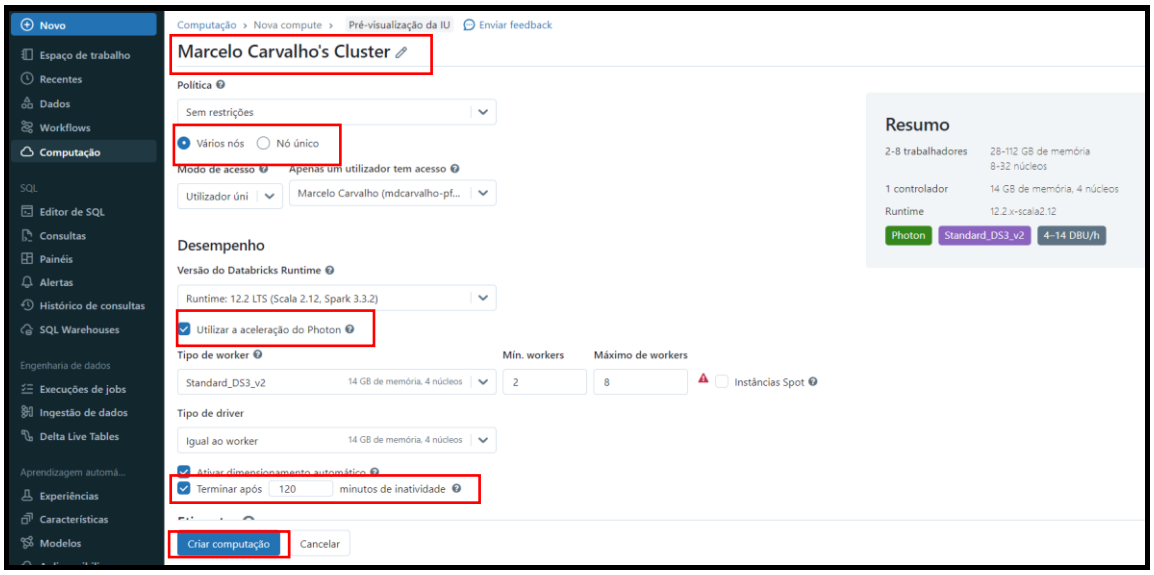


Na tela inicial de “Computação”, clique em “Criar computação”

Na tela de criação do cluster, existem várias configurações disponíveis. Nesse ponto, dependendo do cenário de uso dessa solução (e.g. estudos, experimentações ou ambiente empresarial), talvez seja necessário consultar o time de infraestrutura para definir melhor as capacidades computacionais a serem provisionadas, uma vez que os custos podem variar significativamente.

Para fins de instalação do ambiente com menor custo possível, vamos seguir com as configurações padrão, modificando apenas:

- O nome do cluster (coloque um que faça sentido para você);
- O tipo para nó único;
- Desativar o Photon;
- Reduzir para 10 minutos o tempo de inatividade para término do cluster.

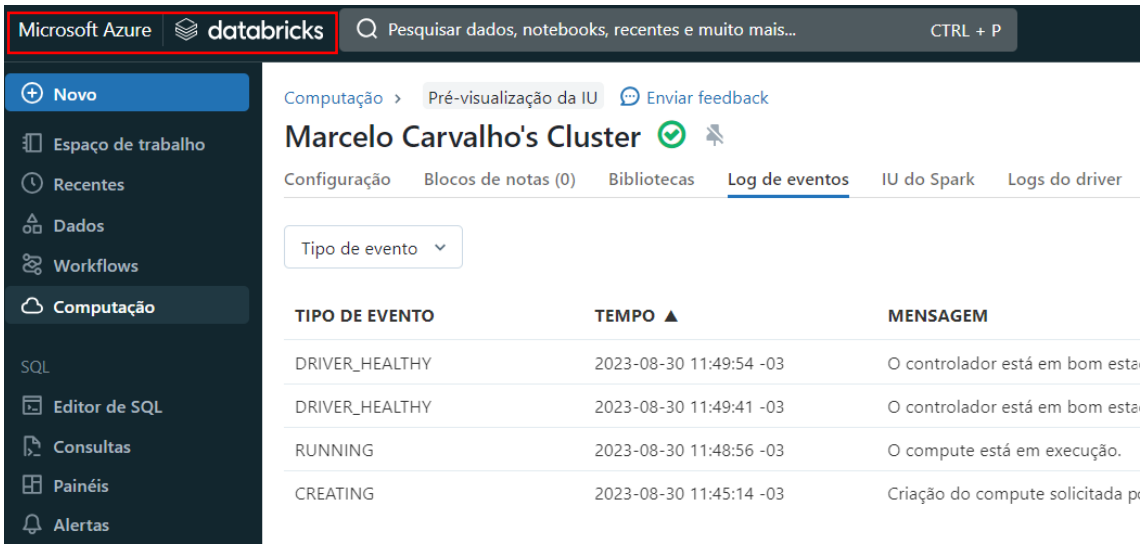


Feitas as configurações, clique em “Criar computação”. Esse processo leva alguns minutos. Você pode acompanhar o progresso clicando em “Log de eventos”.

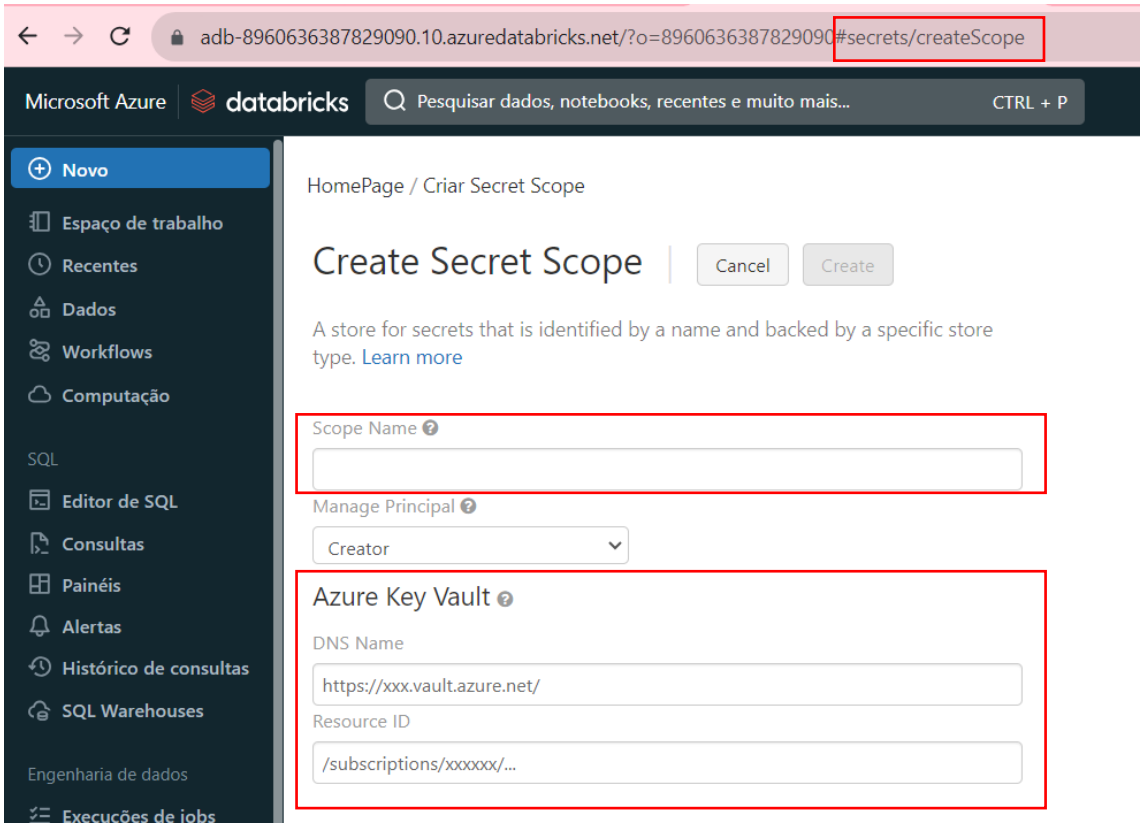
3.3 Secret Scope

O próximo passo da instalação é criar um “secret scope” no Azure Databricks e vincular ao cofre de chaves do Azure.

O primeiro passo é navegar para a tela inicial, clicando em “Microsoft Azure | Databricks”.

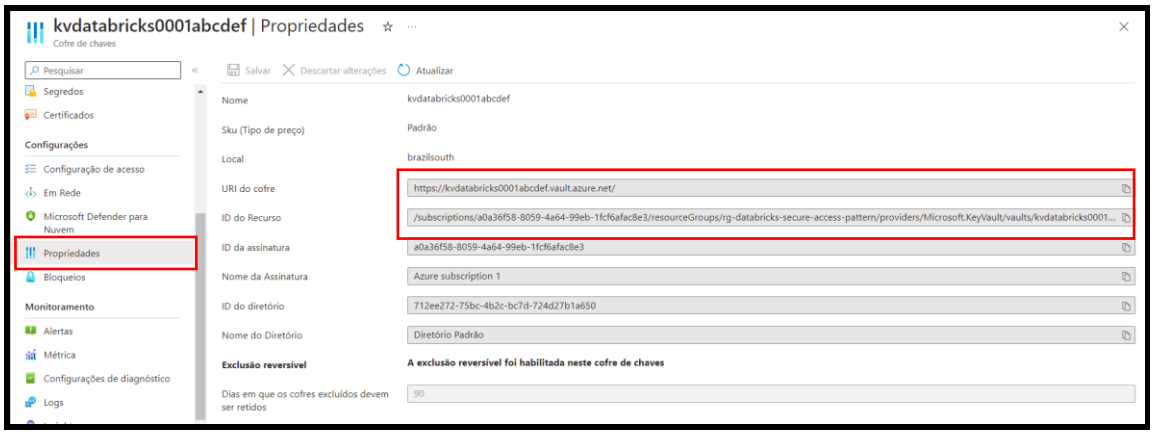


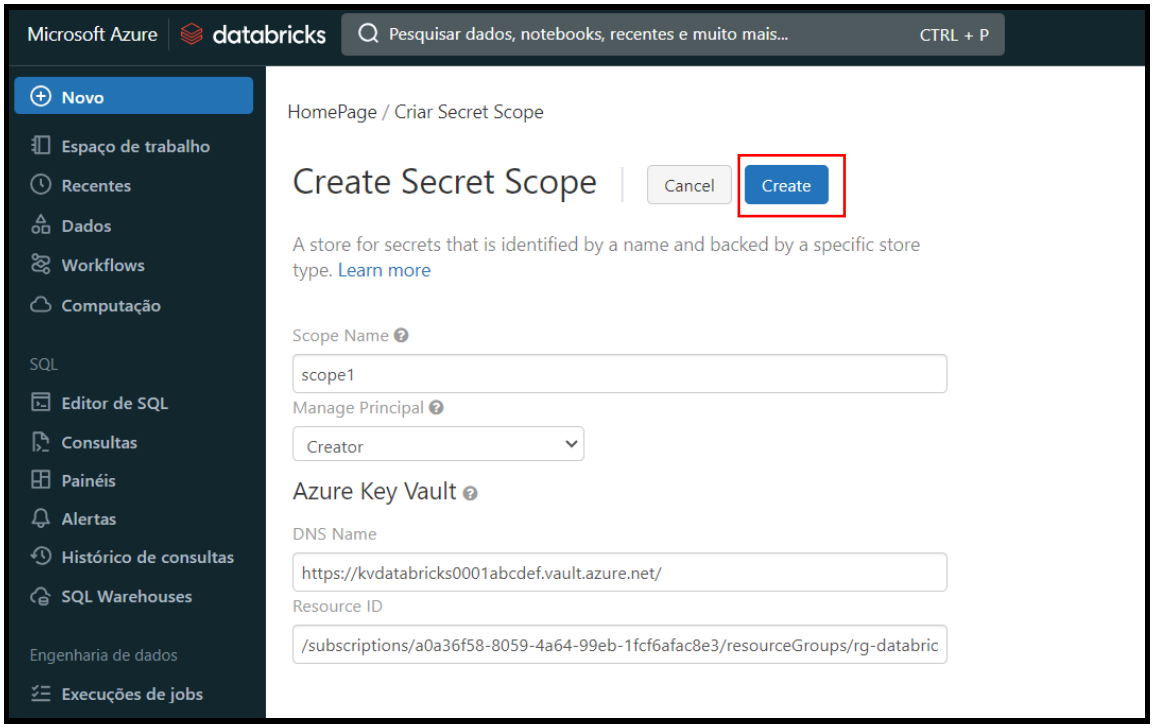
Feito isso, a URL de acesso estará com seu endereço inicial. O “secret scope” fica oculto no Databricks. Para acessá-lo, você precisa complementar a URL ao final com /secrets/createScope



No campo “Scope Name” digite o nome do escopo, à sua escolha. Nos campos do Azure Key Vault você vai preencher com informações do cofre de senhas criado no Azure. Essas informações estão localizadas na tela a seguir.

Volte na tela do cofre de senhas e clique em Propriedades. O valor do campo URI do cofre deve ser copiado e colado no campo DNS Name e o valor do campo ID do Recurso deve ser copiado e colado no campo Resource ID. Após o preenchimento clique em “Create”.

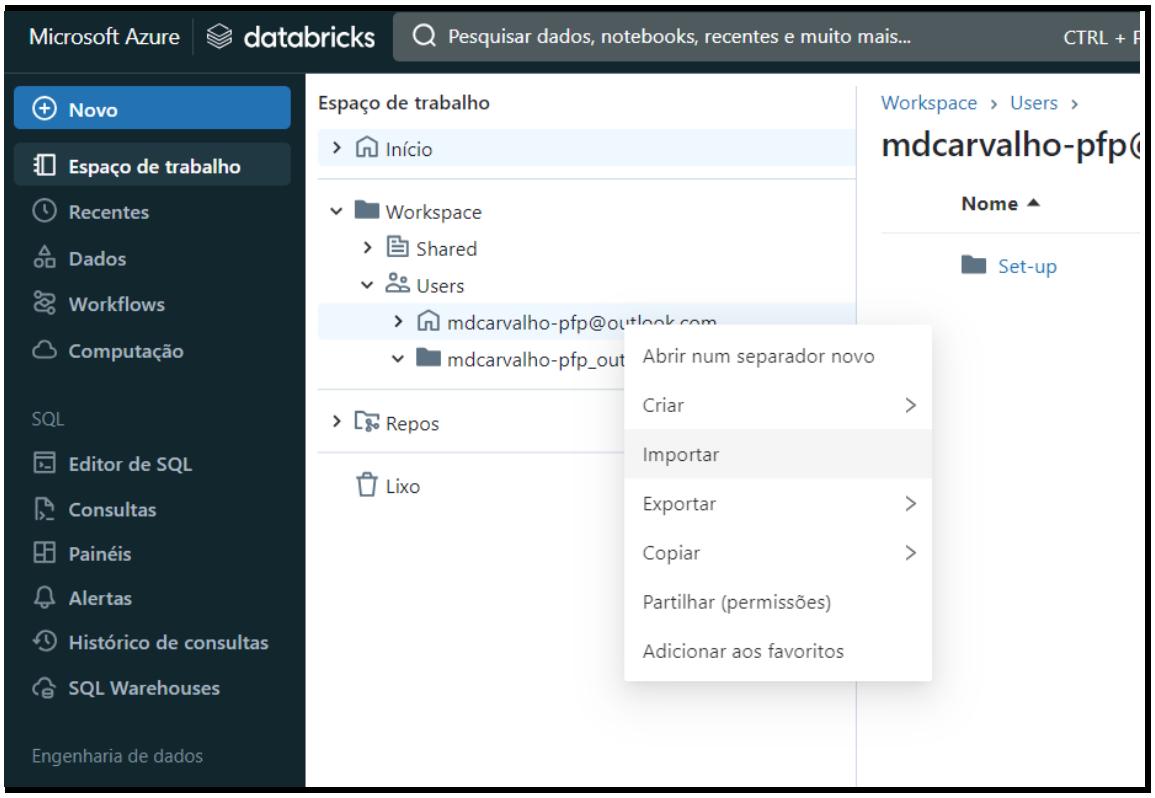




3.4 Teste de acesso ao ADLS a partir do Azure Databricks

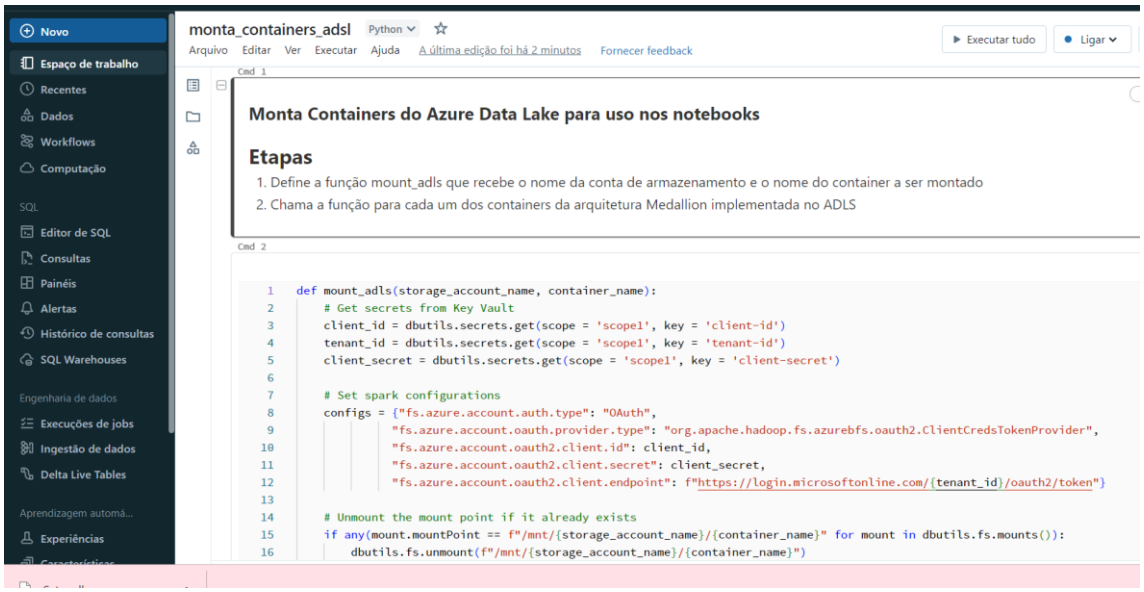
O passo final para verificar se o ambiente está configurado corretamente é executar um notebook no Azure Databricks que monta os containers do ADLS como pontos de montagem no DBFS.

Para esse teste, importe o arquivo setup.dbc no ambiente do Databricks. Para isso clique com o botão direito em seu usuário e escolha “Importar”



Selecione o arquivo setup.dbc na tela de importação

Concluida a importação, a pasta Setup vai aparecer na tela. Clique nela e abra o notebook monta_containers_adsl



Esse notebook é um código em Python que monta os containers do ADLS em pontos de montagem do Databricks para que sejam utilizados nos notebooks. Nas chamadas da função você deve substituir o nome do storage account pelo que você utilizou na sua implantação:

```
mount_adls('<nome da conta de armazenamento', 'bronze')
mount_adls('<nome da conta de armazenamento', 'silver')
mount_adls('<nome da conta de armazenamento', 'gold')
```

E finalmente, clique em “Executar tudo”. Se não tiver mensagem de erro significa que você implantou e configurou o ambiente com sucesso!

3.5 Informações importantes.

De maneira geral, a execução dos passos acima só precisa ser feita uma única vez para um dado locatário (tenant) do Azure. Caso seja necessário executar novamente o script de provisionamento, em função de alguma mensagem de erro, atente para os seguintes pontos:

- Remover o grupo de recursos provisionado.
- Remover o registro de aplicação provisionado.
- Utilizar outros valores de parâmetros informados no template, em relação a execuções anteriores, pois alguns recursos do Azure têm deleção lógica por um certo tempo, impedindo que um novo recurso do mesmo tipo e com o mesmo nome seja provisionado.

4 Glossário

Azure: Microsoft Azure é a plataforma de computação em nuvem da Microsoft. Oferece uma ampla variedade de serviços de computação, armazenamento, banco de dados, análise, rede e outras funcionalidades que organizações podem usar para implantar, gerenciar e dimensionar aplicações.

Azure Databricks: Um serviço Apache Spark baseado em análise colaborativa desenvolvido em colaboração entre Microsoft e Databricks. Facilita a colaboração entre engenheiros de dados, cientistas de dados e analistas de negócios através de notebooks interativos.

Azure Data Lake Storage: Um serviço de armazenamento em escala de petabytes hiperdimensionado para análise de big data. Ele é otimizado para executar análises em grande escala e permite que você execute análises em dados do jeito que quiser (usando, por exemplo, Azure Databricks, HDInsight).

Cofre de Chaves (Key Vault): Azure Key Vault é um serviço de gerenciamento de segredos, chaves e certificados. Ele permite que você armazene e gerencie informações sensíveis de forma centralizada, como chaves de API, senhas e certificados.

Conta de Armazenamento (Storage Account): Um serviço que fornece soluções de armazenamento em nuvem, como blobs, filas, tabelas e discos. Usado para armazenar grandes quantidades de dados, como documentos, backups, informações de log e muito mais.

Registro de Aplicativo (Service Principal): Uma identidade criada para ser usada por aplicativos, serviços e ferramentas automatizadas para acessar recursos específicos do Azure. Ele permite que você defina o escopo e as permissões para a aplicação em um espaço específico do diretório.

Locatário (Tenant): Representa uma organização no Azure AD. É uma instância dedicada e confiável do Azure AD que uma organização recebe no momento em que se inscreve no serviço Microsoft Cloud. O locatário armazena informações sobre os usuários da organização e as informações que eles criam e gerenciam.

PowerShell: É uma linguagem de script baseada em tarefas e um shell de linha de comando desenvolvido pela Microsoft. Ele é usado principalmente para automação e gerenciamento de configuração e é amplamente adotado para administrar sistemas Windows. O Azure oferece o módulo Az para PowerShell, permitindo que os administradores gerenciem os recursos do Azure diretamente a partir da linha de comando do PowerShell.

ARM (Azure Resource Manager): É o serviço de implantação e gerenciamento de recursos do Azure. Ele fornece um modelo de gerenciamento unificado que você pode usar para criar, atualizar e excluir recursos em sua conta Azure. Os templates ARM permitem que os desenvolvedores e administradores de sistemas declarem, em formato

JSON, os recursos que desejam implantar, tornando o processo de implantação altamente repetível e consistente.