

Configurando External Location na AWS:

Crie um bucket S3 que servirá como armazenamento para sua External Location:

The screenshot shows the 'Create bucket' page in the AWS Management Console. The 'General configuration' section is active, showing the 'Bucket name' as 'databrickssgl-external-location' and the 'AWS Region' as 'US West (Oregon) us-west-2'. Below this, there is a section for 'Object Ownership' with two options: 'ACLs disabled (recommended)' (selected) and 'ACLs enabled'. The 'ACLs disabled' option states that all objects in the bucket are owned by this account and access is specified using only policies. The 'ACLs enabled' option states that objects can be owned by other AWS accounts and access is specified using ACLs. At the bottom of the 'Object Ownership' section, it says 'Object Ownership: Bucket owner enforced'.

No AWS IAM, crie uma Role e opte pela opção de Custom Trust Policy:

The screenshot shows the 'Create role' page in the AWS IAM console. The 'Custom trust policy' option is selected. Below this, the 'Custom trust policy' section is visible, showing a JSON policy document. The policy document is as follows:

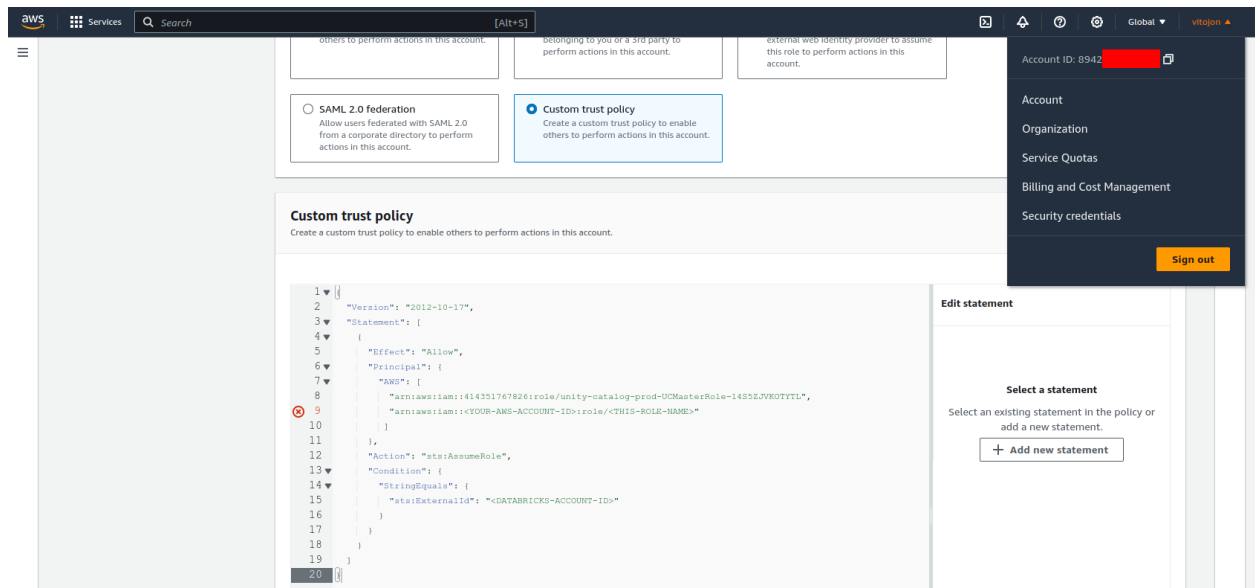
```
1 {
2   "Version": "2012-10-17",
3   "Statement": [
4     {
5       "Effect": "Allow",
6       "Principal": {
7         "AWS": [
8           "arn:aws:iam:414351767826:role/unity-catalog-prod-UCMasterRole-14552JW0TYTL",
9           "arn:aws:iam::<YOUR-AWS-ACCOUNT-ID>:role/<THIS-ROLE-NAME>"
10        ]
11      },
12       "Action": "sts:AssumeRole",
13       "Condition": {
14         "StringEquals": {
15           "sts:ExternalId": "<DATABRICKS-ACCOUNT-ID>"
16         }
17      }
18     ]
19   ]
20 }
```

On the right side of the console, there is a section titled 'Edit statement' with a sub-section 'Select a statement'. It contains the text 'Select an existing statement in the policy or add a new statement.' and a button labeled '+ Add new statement'.

Como já vi o `arn:aws:iam` de UC alterando algumas vezes, vou direcionar o link da documentação (pois pode estar atualizado/alterado na leitura desse PDF) em vez de copiar e colar aqui o JSON da figura:

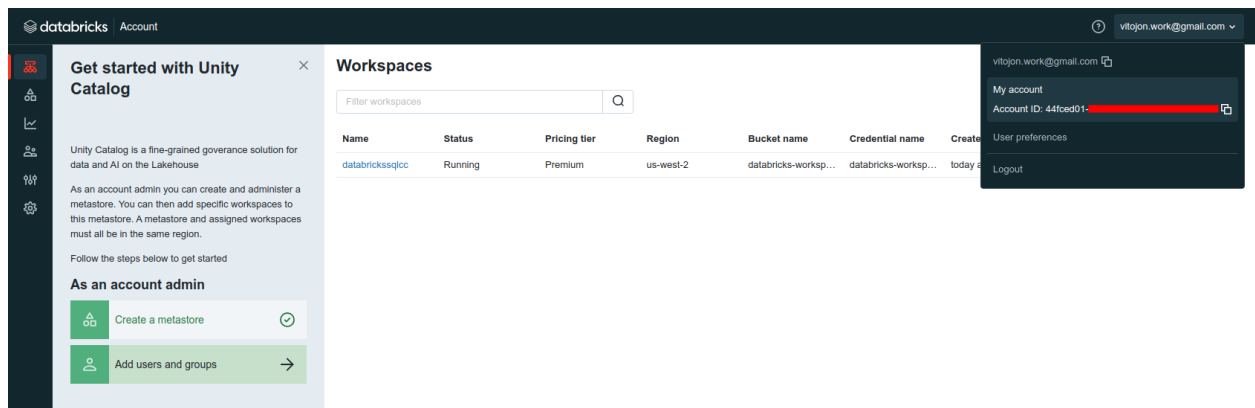
Manage external locations and storage credentials

Conforme visto no JSON da figura anterior, você deve fornecer sua AWS Account ID e Databricks ID, podendo serem obtidas da seguinte maneira respectivamente:



The screenshot shows the AWS IAM console interface. The main content area displays the 'Custom trust policy' configuration for a role. The policy is a JSON document that allows the role to assume the 'unity-catalog-prod-UCMasterRole-14552JYK0TYTL' role in the same account, with a condition that the 'sts:ExternalId' must equal the 'DATABRICKS-ACCOUNT-ID'.

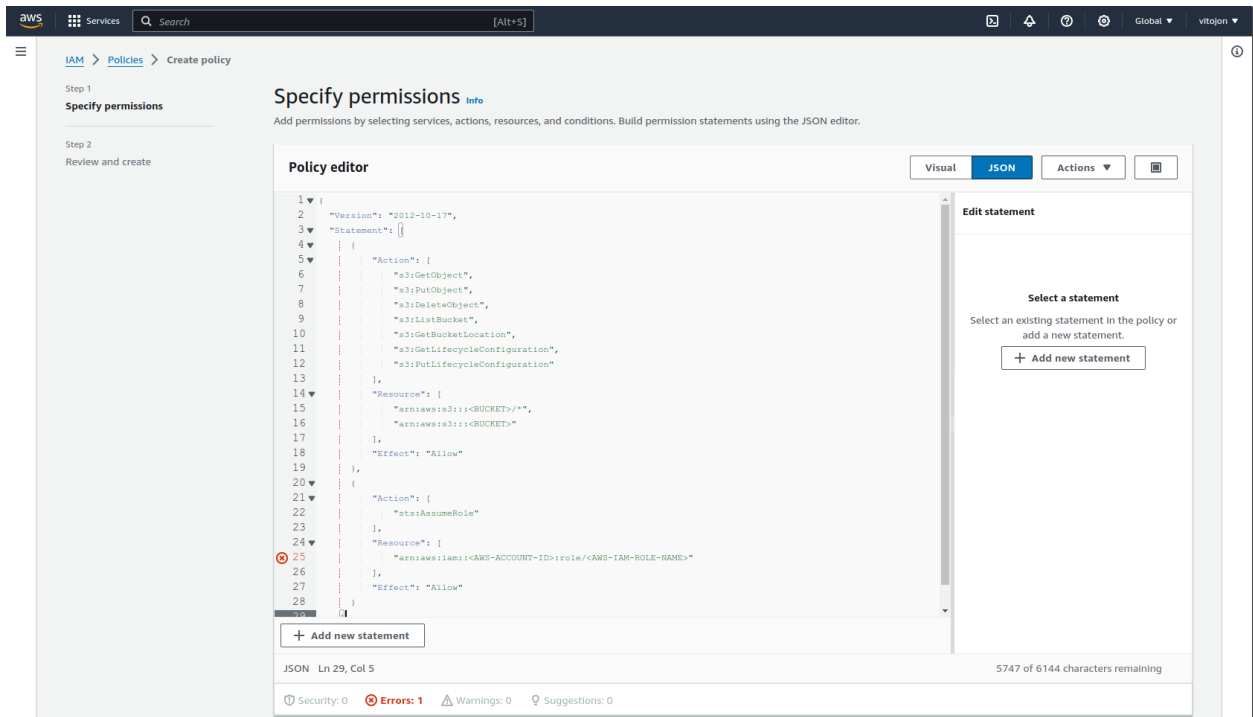
```
1 {
2   "Version": "2012-10-17",
3   "Statement": [
4     {
5       "Effect": "Allow",
6       "Principal": {
7         "AWS": [
8           "arn:aws:iam::414251767826:role/unity-catalog-prod-UCMasterRole-14552JYK0TYTL",
9           "arn:aws:iam::<YOUR-AWS-ACCOUNT-ID>:role/<THIS-ROLE-NAME>"
10        ]
11      },
12      "Action": "sts:AssumeRole",
13      "Condition": {
14        "StringEquals": {
15          "sts:ExternalId": "<DATABRICKS-ACCOUNT-ID>"
16        }
17      }
18    ]
19  }
```



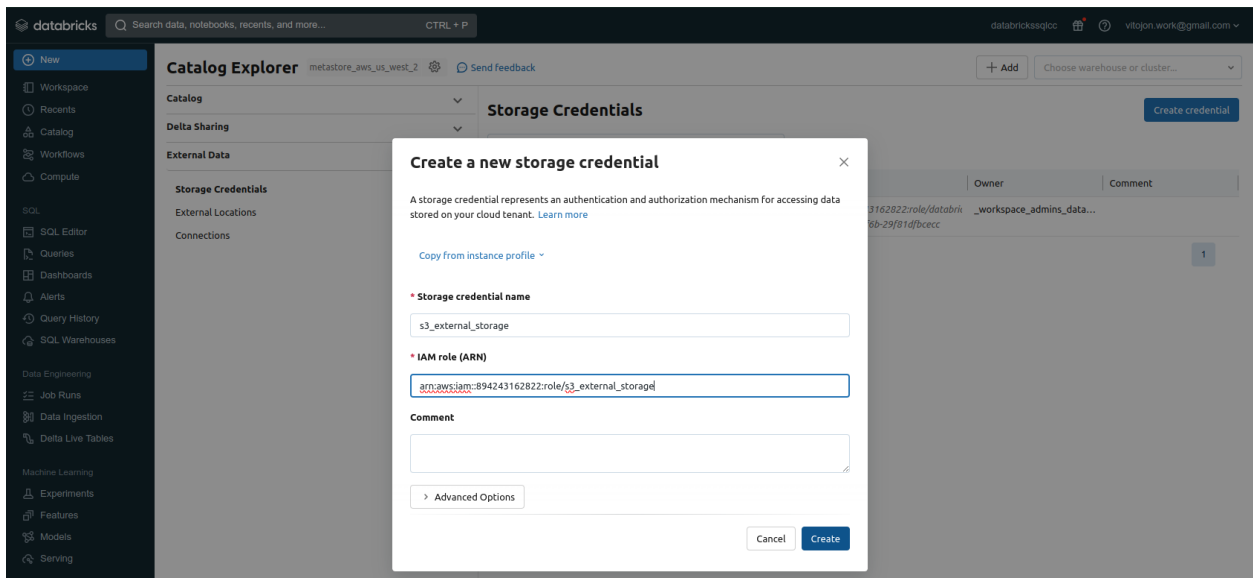
The screenshot shows the Databricks Account console interface. The main content area displays the 'Workspaces' table. The table lists the workspace 'databricksqloc' with status 'Running', pricing tier 'Premium', region 'us-west-2', bucket name 'databricks-worksp...', credential name 'databricks-worksp...', and creation time 'today'.

Name	Status	Pricing tier	Region	Bucket name	Credential name	Create
databricksqloc	Running	Premium	us-west-2	databricks-worksp...	databricks-worksp...	today

Além de Role, é necessário também criar uma Policy, onde também novamente será necessário incluir o AWS ID da conta com sua role de acesso, mas também o bucket que se tornará External Location e foi criado no primeiro passo:



Dentro da Databricks, na página de Catalog, é possível visualizar um espaço direcionado para External Data, vá em Storage Credentials para criar uma credencial relacionada à role configurada no IAM:



E uma vez existente a Storage Credential, a AWS irá te disponibilizar um ID dessa conexão:

Storage credential created

Please navigate to the AWS console and configure your IAM role with the below value as the externalID to establish a cross account trust relationship. [Learn more](#)

IAM role (ARN):
arn:aws:iam::894243162822:role/s3_external_storage

External ID

44fced01-8b32-4e0a-9f6b-29f81dfbcecc

Done

Se direcione até o espaço de criação de External Locations (logo abaixo ao Storage Credentials) e lá você terá a opção Quickstart (similar à criação de Workspace/Unity Catalog via template da Stack no AWS CloudFormation):

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Experiments

Catalog Explorer

metastore_aws_us_west_2

Send feedback

+ Add

Choose warehouse or cluster...

Catalog

Delta Sharing

External Data

Storage Credentials

External Locations

Connections

External Locations

Filter locations

1 location

Name	Credential	URL	Owner	Comment
databricks-external-locations-63120-bucket/uni...	databricks-external-locations-63120-bucket/uni...	s3://databricks-external-locations-63120-bucket/uni...	workspace_admins_databricksq...	

Create location

Create a new external location

An external location is a cloud storage url (and paired credential) that allows access to data stored on your cloud tenant. [Learn more](#)

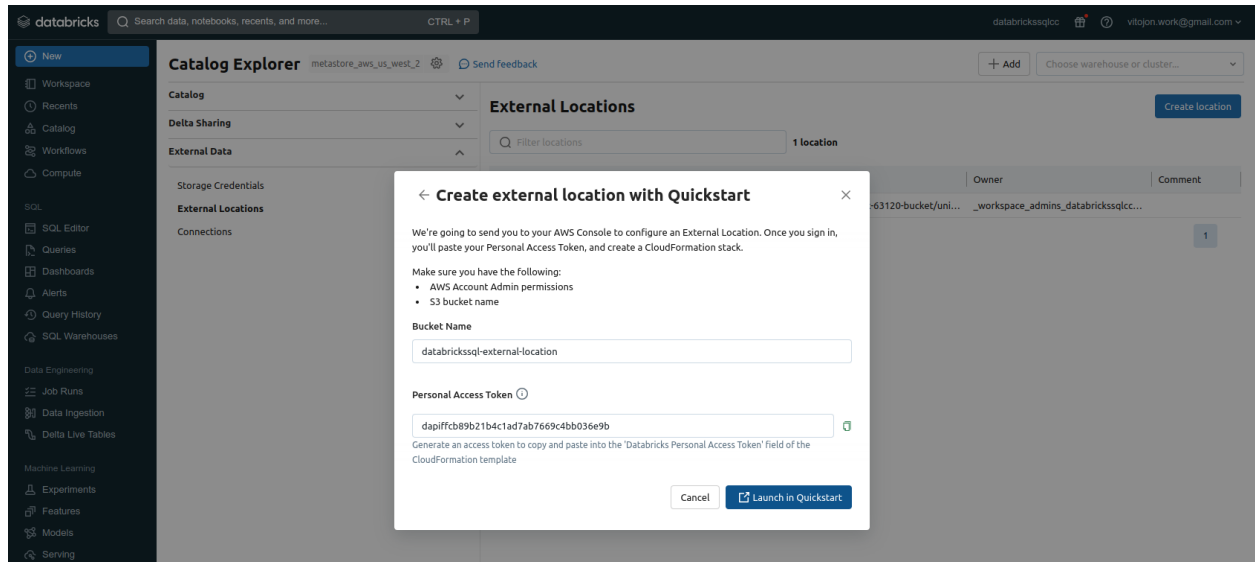
How would you like to create an external location?

☒ AWS Quickstart (Recommended)
Use Quickstart to create an S3-based external location in a few clicks

☐ Manual
For advanced users, users who already have a storage credential for use with a bucket, or users who want to use cloud storage other than S3

Cancel Next

Para que seja iniciado o CloudFormation, forneça o bucket S3 que foi criado no primeiro passo e então submetido ao permissionamento em passos seguintes. Por fim, clique na região do Personal Access Token para gerar um Databricks token que irá habilitar capacidades para que o CloudFormation template funcione e tenha acesso para gerar External Location:



Insira o Personal Access Token gerado como um dos parâmetros no template do CloudFormation:

Parameters
Parameters are defined in your template and allow you to input custom values when you create or update a stack.

Databricks Account Credentials

Databricks Personal Access Token
A personal access token for your Databricks account.

.....

Databricks account ID
Find your account ID at <https://accounts.cloud.databricks.com>

44fced01-8b32-4e0a-9f6b-29f81dfbcecc

Workspace configuration

URL of your Databricks workspace (e.g. <https://your-workspace.databricks.com>)
Url of your workspace.

<https://dbc-f129c9cf-275e.cloud.databricks.com>

Data bucket name
The S3 bucket where your data is stored.

aws-external-storage

Databricks UC Role ARN (DO NOT MODIFY)
This is a static value that references a role created by Databricks

arn:aws:iam::414351767826:role/unity-catalog-prod-UCMasterRole-1455ZJVKOTYTL

E por fim aguarde a criação da Stack (acesso e comandos associados ao External Location se encontram em notebook do Dia 4):

The screenshot displays the AWS CloudFormation console interface. On the left, the 'Stacks' list shows two stacks: 'databricks-s3-ingest-d076c' (status: CREATE_IN_PROGRESS) and 'databricks-workspace-stack-a8813' (status: CREATE_COMPLETE). The main panel shows the details for the 'databricks-s3-ingest-d076c' stack, including tabs for Stack info, Events, Resources, Outputs, Parameters, Template, Change sets, and Git sync. The 'Events' tab is active, showing a single event with the status 'CREATE_IN_PROGRESS' and the reason 'User initiated'.

Stacks (2)

Filter status: Active

View nested

Stacks

- databricks-s3-ingest-d076c**
2023-11-28 11:09:14 UTC-0300
CREATE_IN_PROGRESS
- databricks-workspace-stack-a8813
2023-11-28 10:41:56 UTC-0300
CREATE_COMPLETE

databricks-s3-ingest-d076c

Stack info | **Events** | Resources | Outputs | Parameters | Template | Change sets | Git sync - new

Events (1)

Search events

Timestamp	Logical ID	Status	Status reason
2023-11-28 11:09:14 UTC-0300	databricks-s3-ingest-d076c	CREATE_IN_PROGRESS	User initiated

Configurando External Location na GCP:

Crie seu bucket GCS que servirá como armazenamento do External Location:

The screenshot shows the 'Create a bucket' wizard in the Google Cloud Platform console. The first step, 'Name your bucket', is completed with the name 'gcp_external_location_bucket_for_databricks'. The second step, 'Choose where to store your data', is active. It offers three location types: 'Multi-region' (highest availability), 'Dual-region' (high availability and low latency), and 'Region' (lowest latency). The 'Region' option is selected, and a dropdown menu shows 'us-central1 (Iowa)' as the chosen location. A 'CONTINUE' button is at the bottom. To the right, a 'Good to know' section provides information on location pricing, showing a current configuration of 'Region / Standard' with a cost of '\$0.020 per GB-month' for 'us-central1 (Iowa)'. An 'ESTIMATE YOUR MONTHLY COST' link is also present.

Na plataforma da Databricks, nomeie uma credential a ser utilizada e receba um Service Account após a criação:

The screenshot displays the 'Storage Credentials' section in the Databricks Catalog Explorer. A table lists two credentials: 'e68472b0-3acb-4403-9eab-d3...' and 'gcp_credential', both of type 'Service Account'. A 'Create credential' button is in the top right. A modal dialog titled 'Storage credential created' is open, displaying the email 'db-uc-storage-06614ulgas-xe0dy@uc-uscentral1.iam.gserviceaccount.com' and a 'Done' button. The modal also includes instructions to grant 'Storage Legacy Bucket Reader' and 'Storage Object Admin' roles to the service account.

Name	Credential Type	Properties	Owner	Comment
e68472b0-3acb-4403-9eab-d3...	Service Account	Email: db-uc-storage-06614tdsf4-lvimi@uc-uscentral1.iam.gserviceaccount.com	vitojon.work@gmail.com	
gcp_credential	Service Account	Email: db-uc-storage-06614ulgas-xe0dy@uc-uscentral1.iam.gserviceaccount.com	vitojon.work@gmail.com	

Serão necessárias as mesmas permissões que existem na criação de Unity Catalog (“Storage Legacy Bucket Reader” e “Storage Object Admin”), exigindo então uma configuração nas permissões do bucket:

The screenshot shows the Google Cloud IAM console for a bucket named 'gcp_external_location_bucket_for_databricks'. The 'Permissions' section is active, showing a list of principals and their roles. The 'Add principals' section is also visible, showing the role 'Storage Legacy Bucket Reader' and 'Storage Object Admin' being assigned to the principal 'db-uc-storage-06614ulgas-xe0dy@uc-uscentral1.iam.gserviceaccount.com'.

Permissions

Type	Principal	Name	Role
Service Account	db-4684743289196049@prod-gcp-us-central1.iam.gserviceaccount.com		Databricks
Group	Editors of project: weighty-skyline-404619		Storage Legacy Bucket Reader
Group	Owners of project: weighty-skyline-404619		Storage Legacy Bucket Admin
Service Account	service-283963431256@compute-system.iam.gserviceaccount.com	Compute Engine Service Agent	Compute Engine
Service Account	service-283963431256@containerregistry.iam.gserviceaccount.com	Google Container Registry Service Agent	Container Registry
Group	Viewers of project: weighty-skyline-404619		Storage Legacy Bucket Reader

Add principals

Principals are users, groups, domains, or service accounts. [Learn more about principals in IAM](#)

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

Role *

Storage Legacy Bucket Reader

IAM condition (optional)

+ ADD IAM CONDITION

Role *

Storage Object Admin

IAM condition (optional)

+ ADD IAM CONDITION

+ ADD ANOTHER ROLE

SAVE **CANCEL**

Uma vez dado o permissionamento no bucket à Service Account fornecida, é possível então ir até a tela da Databricks para criação do External Location (acesso e comandos associados ao External Location se encontram em notebook do Dia 4):

The screenshot shows the Databricks Catalog Explorer interface. A dialog box titled 'Create a new external location' is open, allowing the user to configure a new external location. The dialog includes fields for 'External location name', 'Storage credential', 'URL', and a 'Comment' field. The 'Storage credential' dropdown is set to 'gcp_credential (Service Account)', and the 'URL' field contains 'gs://gcp_external_location_bucket_for_databricks'.

Create a new external location

An external location is a cloud storage url (and paired credential) that allows access to data stored on your cloud tenant. [Learn more](#)

*** External location name**

gcs_external_location

*** Storage credential**

gcp_credential (Service Account)

Email: db-uc-storage-06614ulgas-xe0dy@uc-uscentral1.iam.gserviceaccount.com

*** URL**

gs://gcp_external_location_bucket_for_databricks

Comment

> Advanced Options

Cancel **Create**

Configurando External Location na Azure:

Assim como na criação do Unity Catalog, crie um bucket na Azure com performance Premium e de tipo Block Blobs:

Microsoft Azure

Search resources, services, and docs (G+)

Home > Storage accounts >

Create a storage account

Basics | Advanced | Networking | Data protection | Encryption | Tags | Review

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * Azure subscription 1

Resource group * databricks_workshop_mdw

Instance details

Storage account name * azureexternalstorage

Region * (US) East US 2

Performance * ☐ Standard: Recommended for most scenarios (general-purpose v2 account) ☒ Premium: Recommended for scenarios that require low latency.

Premium account type * Block blobs

Redundancy * Locally-redundant storage (LRS)

Após criação bem sucedida do seu Storage, crie o seu Container:

azureexternalstorage | Containers

Storage account

Search

+ Container | Change access level | Restore containers | Refresh | Delete | Give feedback

Search containers by prefix

Name	Last modified	Anonymous access level
<input type="checkbox"/> slogs	11/27/2023, 8:46:35 PM	Private

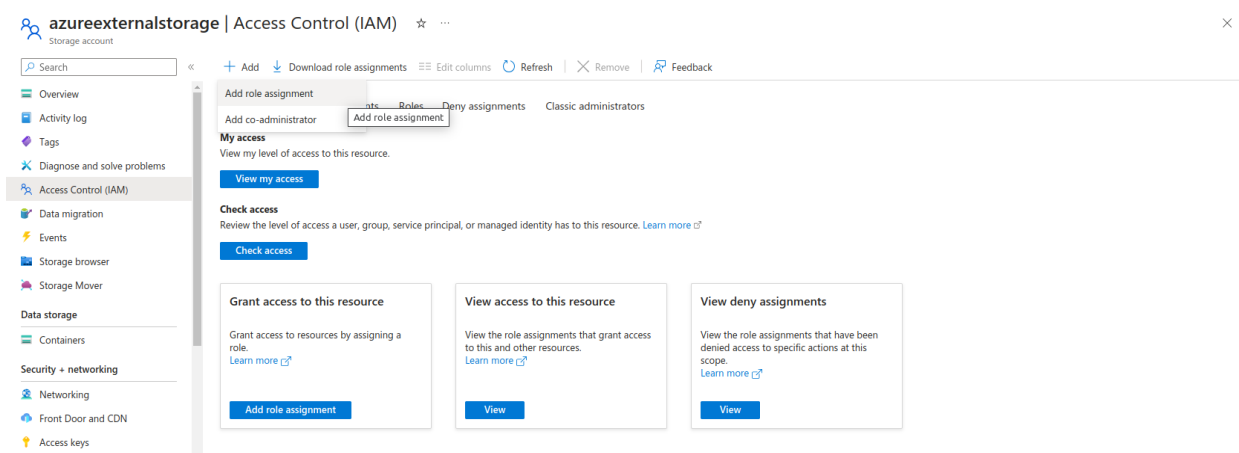
Anonymous access level * externalstorage

Anonymous access level Private (no anonymous access)

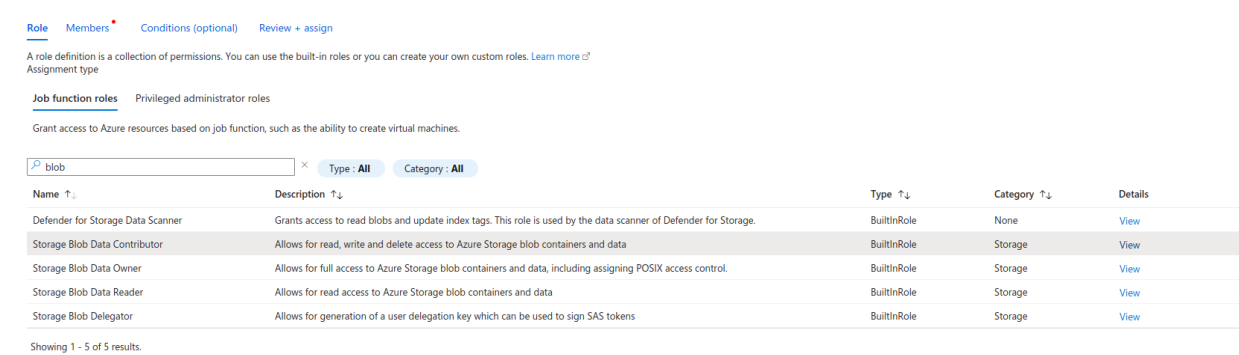
The access level is set to private because anonymous access is disabled on this storage account.

Advanced

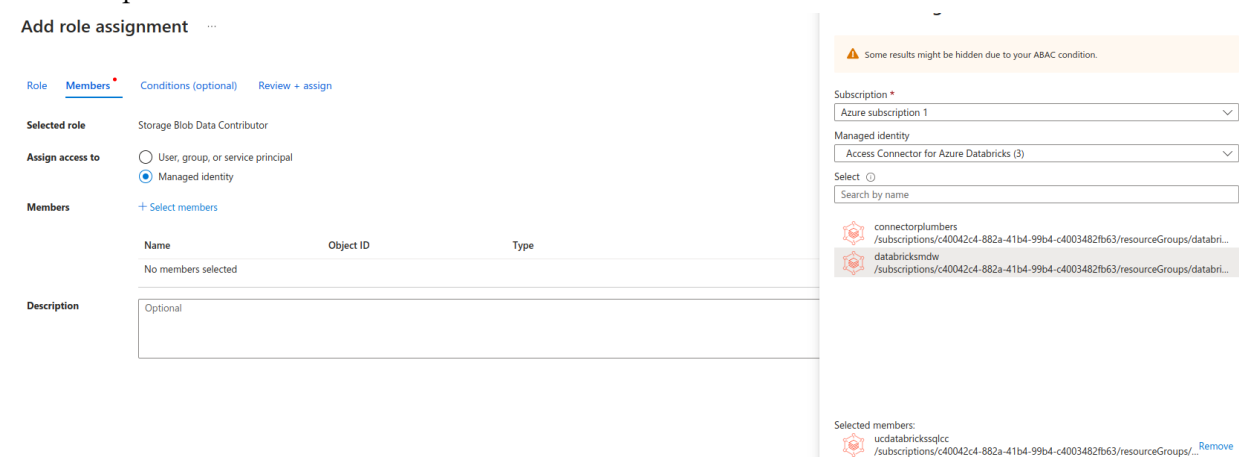
Se dirija até a IAM do seu Storage e adicione uma role para que o conector possa interligar o Gen 2 com a Databricks, assim como feito com Unity Catalog, podendo inclusive reaproveitar o mesmo conector em caso a depender de Resource Group e região:



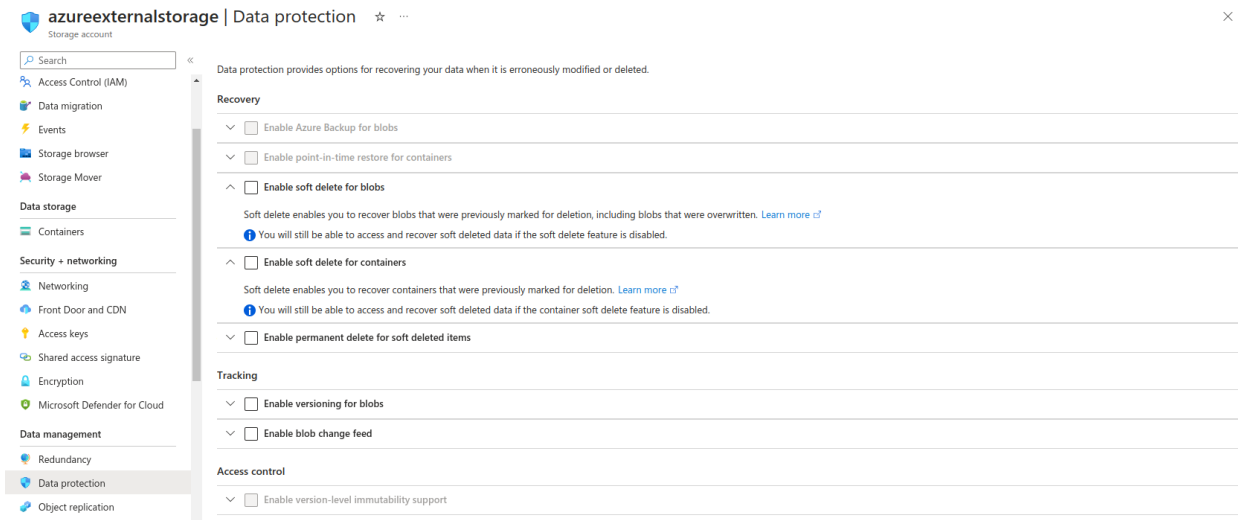
Você irá incluir a opção de **Storage Blob Data Contributor** para o conector criado na aba de Roles:



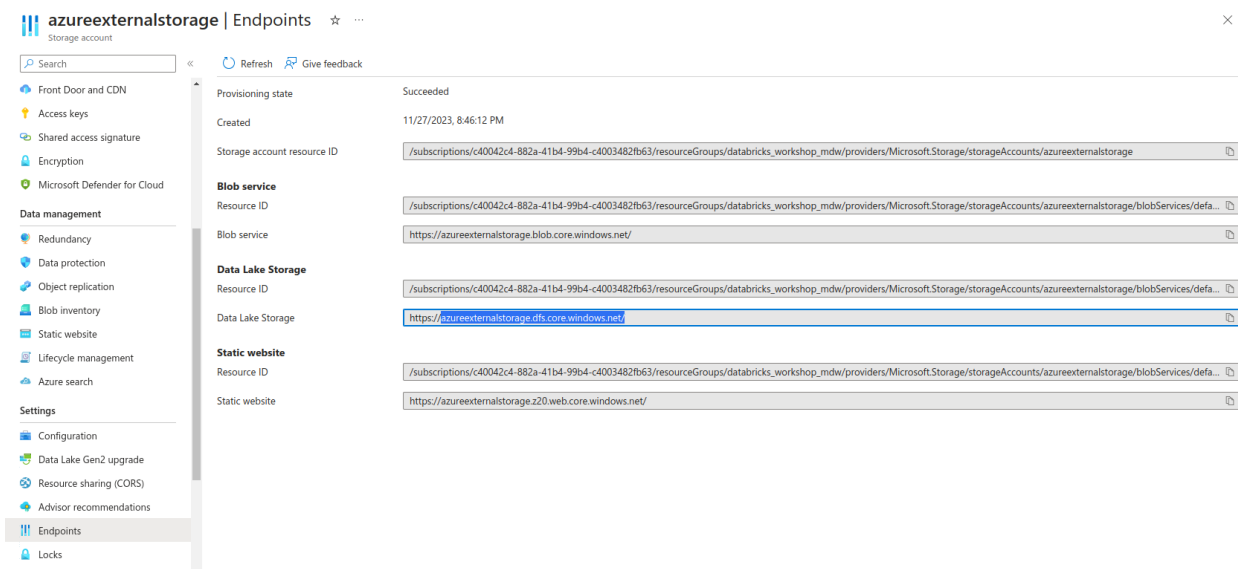
Já na aba de Members, adicione uma Managed Identity, onde poderá ser encontrado o seu conector para Databricks:



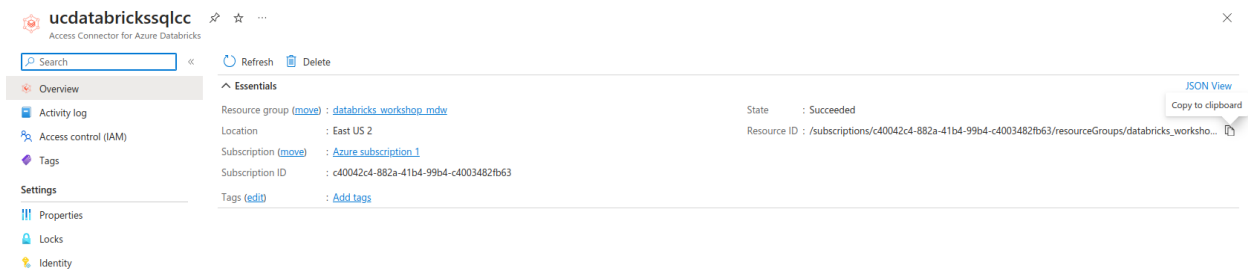
Ajuste as configurações de Soft Delete nas configurações de management de Data Protection, removendo as opções de habilitar soft delete em Blobs e Containers:



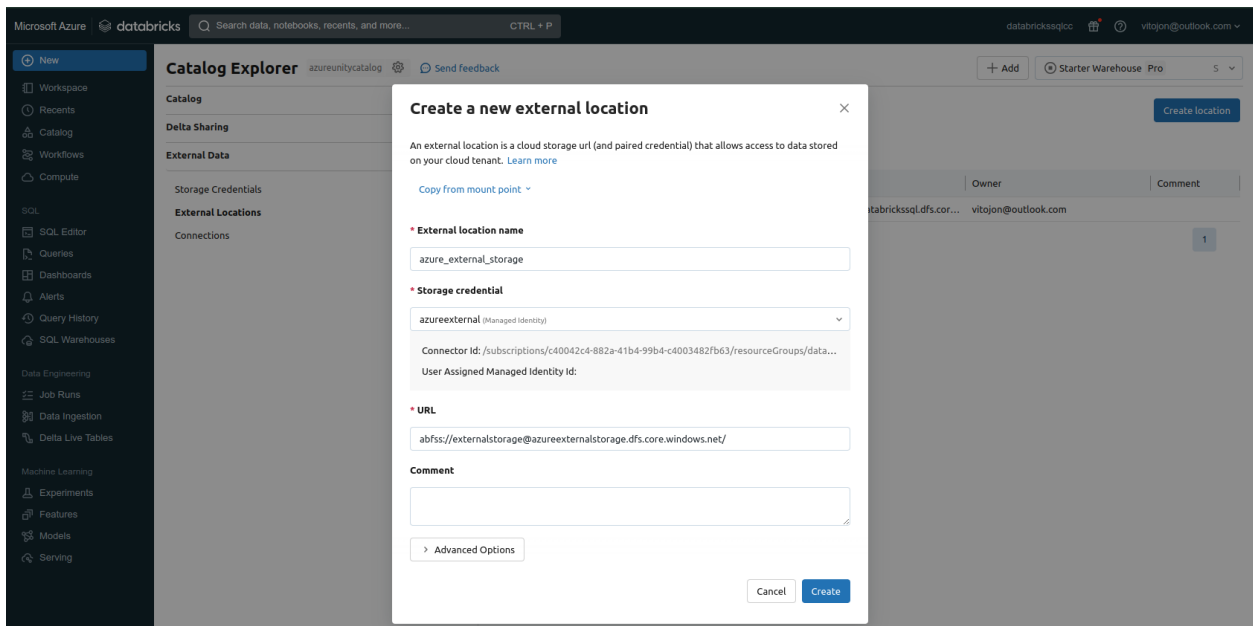
Por fim, se preferir, pode ir até o campo de Endpoints no Storage para copiar seu endereço de Storage (será pedido na Databricks):



Assim como também deve copiar o Resource ID do seu conector:



Compareça na página de Metastore da Account Console em Data:



Region: mesma que do Workspace

ADLS Gen 2 Path formato: <container_name>@<storage_account_name>.dfs.core.windows.net
(possível ser copiado o trecho em Endpoints precisando adicionar apenas o container@)

Access Connector Id: copiar resource ID da página do seu conector.

