Azure Data Governance

Azure (UC + RBAC)

P2 Licenced subscription is needed for RBAC group creation
UC and RBAC is accessed through AAD

RBAC originally from azure but UC originally from DB

PERVIEW Liniage	**vs**	**UC Liniage**
Perview is global		UC is limited to azure enviroment
Preview is Azure Service		UC is used with AWS / Azure / GCP
Perview Connects with on prem		UC is limited to Azure enviroment only
Preview Scans file in onprem and rest of the azure env		UC only Scans files in azure env

Preview Scans file on onprem and rest of the azure env – Suppose someone wants to get some data from any point in pipeline, may it be onprem or files located in any storage layers, Perview can do it. It can collect this data right from onprem and azure platform and display it in a dashboard. This can be beneficial to understand the underlying cost of the services used

We create RBAC group under microsoft Entra ID
Creating ADD groups and adding mail id's for access is not a job of DE

SCIM Connector is an in build ADB application which is used to sync the AAD groups from AZ Entra ID to the databricks account portal.

Flow – Reflection of AAD groups in DB Workspace

1. Creation of AAD env (with P2 Subscription)
2. Creation of AAD groups
3. Adding users to AAD
4. Creation of SCIM Connector (and connect it with ADB using tokens and tenate ID (get them from ADB))
5. Add AAD groups in SCIM connectors (here after 40 Min, groups will reflect in ADB acccount group level)
6. Verify these AAD groups are SYNCED with Databricks account
7. Add groups from ADB acccount group to **ADB workspace**

If you have 40 groups and you want those to reflect it in ADB account groups level, it will take minimum 40 min. You can do this manually as well where one group will be added to ADB account groups level one at a time. This one at a time approch is usefull if you have some add hoc requirement for a group that should be reflected in ADB account groups level immediately.

Account Admins Vs Metastore Admins Vs Workspace Admins

DBX Account Admin has complete access of all workspaces across enviroment

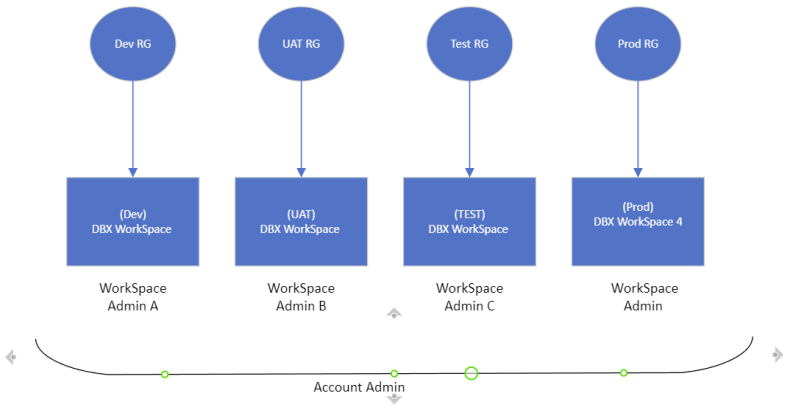
Users – Limited access to resources in ADB workspace
WorkSpace Admin – as Admin you have complete access to a **Specific (Dev / UT / Prod)** ADB workspace.

Also, as workSpace admin you have complete access of MetaStore of that Specific ADB WorkSpace

UC – Datagovernance

Hierarchy in Azure databricks

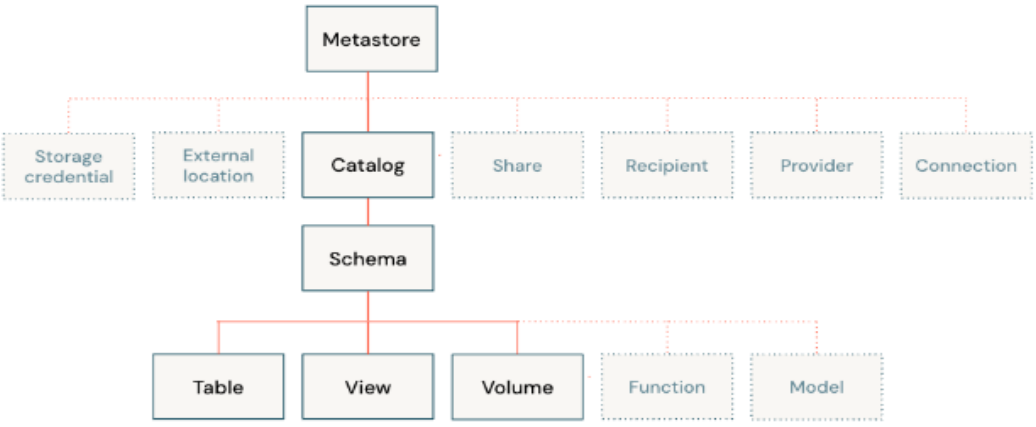
Role	Description
Globlal Admin	Complete access on Azure account
Metastore Admin	Complete access on databricks account portal & has admin privileges to metatsore
Workpsace Admin	Complete access on a particular Databricks workspace
Unity Catalog owner	has all privileges on a catalog along with assign roles to AAD groups on that catalog
Schema owner	has all privileges on a schema along with assign roles to AAD groups on that schema
Table owner	has all privileges on a Table along with assign roles to AAD groups on that Table
Storage credentials owner	has all privileges on a stoarge credential along with assign roles to AAD groups on that stoarge credential
External location owner	has all privileges on a external location along with assign roles to AAD groups on that external location



Here Account Admin can also be called as Infa Admin

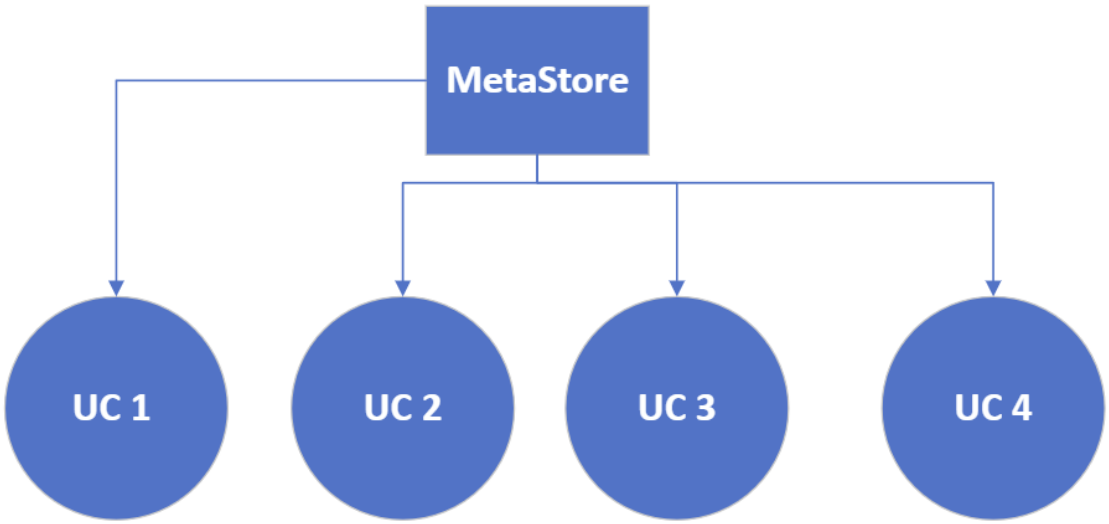
Metastore

A metastore is the top-level container of objects in Unity Catalog. It registers metadata about data assets and the permissions that govern access to them



Number of # that can be created under 1 Meta Store
We can inly create 1 MetaSore per region

Object	Parent	Value
table	schema	10000
table	metastore	100000
volume	schema	10000
function	schema	10000
registered model	schema	1000
registered model	metastore	5000
model version	registered model	10000
model version	metastore	100000
schema	catalog	10000
catalog	metastore	1000
connection	metastore	1000
storage credential	metastore	200
external location	metastore	500



UC – 1 Per Project
We use only one UC per project
Whenever you create a workSpace, UC and Metastore get automatically created
We will use UC created when we create dev DB enviroment across all the enviroment and manage the access of that UC with other env via ADD

Structure of UC:

Previously in case of Hive metastore :

Project 1

Dev DB Env :

- Dev raw Schema
- Dev Intermediate Schema
- Dev Curated Schema

Test DB Env :

- Test raw Schema
- Test Intermediate Schema
- Test Curated Schema

UAT DB Env :

- UAT raw Schema
- UAT Intermediate Schema
- UAT Curated Schema

Prod DB Env :

- Prod raw Schema
- Prod Intermediate Schema
- Prod Curated Schema

After UC came into Play

UC:

- Dev raw Schema
- Dev Intermediate Schema
- Dev Curated Schema
- Test raw Schema
- Test Intermediate Schema
- Test Curated Schema
- UAT raw Schema
- UAT Intermediate Schema
- UAT Curated Schema
- Prod raw Schema
- Prod Intermediate Schema
- Prod Curated Schema

In order to sync from hive metaStore to UC, the cluster needs to be UC enabled

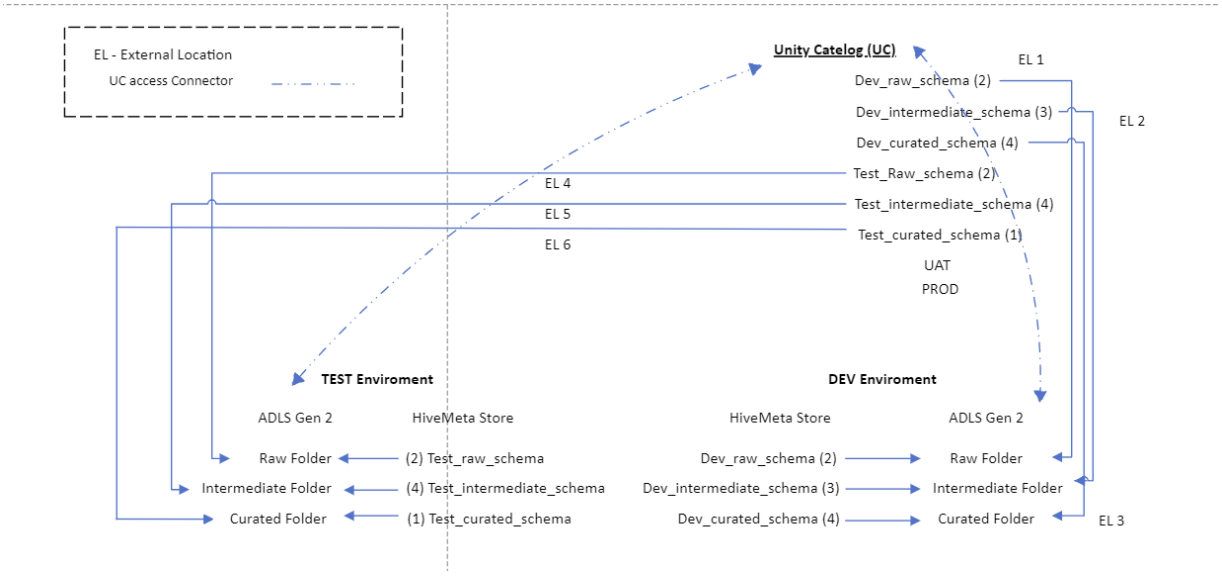
Flow – Reflection of AAD groups in DB Workspace
(Steps marked in green are already done)

1. Creation of AAD env (with P2 Subscription)
2. Creation of AAD groups
3. Adding users to AAD
4. Creation of SCIM Connector (and connect it with ADB using tokens and tenate ID (get them from ADB))
5. Add AAD groups in SCIM connectors (here after 40 Min, groups will reflect in ADB acccount group level)
6. Verify these AAD groups are SYNCED with Databricks account
7. Add groups from ADB acccount group to **ADB workspace**

8. Create managed schemas under Unity Catalog
    a. You can do this using a simple notebook

9. Create storage credentials in databricks workspace
    a. When you create databriks workspace, UC and UC access conenctor gets automatically created.
    b. You can find UC access connector in the default RG created by databricks when you first create databricks workspace.
    c. To create storage credentials in DB workspace, there are two steps.
        i. Give access of ADLS gen 2 to UC Access connector using IAM
        ii. Put access connetor ID into storage credentials

10. Create external locations
    a. create external location for each container in storage account



    b.

11. SYNC schemas from Hive metastore to Unity Catalog
    a. Use in built SYNC commands for migration
12. Assign necessary roles to AAD groups

AAD Group name	role
AAD_METASTORE_ADMIN	Metastore Admin
AAD_TELECOM_DBXWS_ADMIN_DEV	Workspace admin for dev
AAD_TELECOM_DBXWS_ADMIN_PROD	Workspace admin for prod
AAD_TELECOM_DE	DE Team + UC owners + All prvilleges on hive metastore
AAD_TELECOM_REPORT	Reporting Team- Access on curated tables
AAD_TELECOM_TESTING	Testing team- Read access on Unity Catalog,Schemas, Tables/Views
AAD_INFRA_ADMIN	Infra admin team- All previlleges on all databricks envs

This above process is basically **migration of data** from Hivemetastore to UC

When you sync the schemas from hive metastore to UC, only tables would come to UC if you don't have storage credentials lodged. So, to sync data as well as table, you need to authenticate storage credentials in the catelog section

To authenticate external storage, we have storage credentials.

If you don't create external location, in that case DB will use default external location and create the tables in its default storage location as a managed table. This default storage location is created when we create DB workSpace for the first time.

UC access connector comes into play only when you want to sync the Hive MetaSore with UC.
If you want to create table in UC directy, the SPN connection will work in the same way as it worked in case of table creation of Hive Metastore. In that case we don't need to create external location itself therefore UC access connector won't come into play. Sprint 7 / Part 5 / 22:32

Data Lineage : Life cycle of the data
Helps determine the history of the data ultimately understanding its origin, data Quality and data regulatory compliance. All of this addups to having more trust in your data

Where are data comes
Validating accuracy and consistency of the data
What kind of transformation are done on the data to help you achieve regulatoy complaince

Use UC to run the script and you should be able to see the data lineage of the pipeline.
data lineage will show you info of latest pipeline run

Unity catelog real wold Example

Imagine a multinational corporation that uses Azure Databricks for predictive maintenance of its manufacturing equipment. The engineering team is responsible for creating machine learning models, the IT team manages the data infrastructure, and the compliance team ensures that all data activities meet regulatory standards. Unity Catalog serves as the common platform where all these teams can come together. The engineering team can easily find the datasets they need for model training, the IT team can manage permissions and monitor data usage, and the compliance team can keep track of data lineage and access logs. This cohesive approach streamlines operations and ensures better data governance.

Unity Catalog is not just an add-on but a fundamental component of the Azure Databricks ecosystem, designed to meet the growing needs for data governance and security in today's complex data landscape