Databricks SQL for Data Analytics Manual

Unity Catalog

A metastore can contain a number of different catalogs. Within catalogs you can have schemas, and within a schema, you can have a range of tables and views.

A screenshot of a computer

Description automatically generated

Databrick before eunity catalog

A screenshot of a web page

Description automatically generated

Each workspace used to have its own hive metastore. This is a legacy metastore, and should not be confused with the unity catalog. It used to store metadata for the workspace, and user access and based on the individual workspaces.

With Unity Catalog

The unity catalog takes care of user group management, and access control and there is a unity catalog metastore. A single metastore can be assigned to multiple workspaces. There is one unity metastore per region. Each unity catalog is connected to an adls container in azure. Each workspace still ha

A screenshot of a computer

Description automatically generated

Unity catalog benefits

A close-up of a document

Description automatically generated

A close-up of a data catalog

Description automatically generated

Unity catalog requirements

A close-up of a catalog

Description automatically generated

Access connector for Azure Databricks

This is a connector that allows databricks to securely access data that is stored in azure data lake storage. It provides seamless integration between the two.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

After creating, this will now appear below:

A screenshot of a computer

Description automatically generated

You now want to assign adls access to the access connector, this can be done in adls IAM. Add a role assignment as storage blob data contributer.

A screenshot of a computer

Description automatically generated

Assign access to the managed identity ‘access connector for azure databricks’

A screenshot of a computer

Description automatically generated

Check if you are a global administrator in entra id.

A screenshot of a computer

Description automatically generated

Creating a databricks workspace

For unity to work, you need to create premium workspaces as below.

A screenshot of a computer

Description automatically generated

Enabling Unity Catalog

A screenshot of a computer

Description automatically generated

Go to manage account in the top right within the databricks premium workspace. Then go to menu options and expand, and then click on Data.

A screenshot of a computer

Description automatically generated

Lets first create a new container in adls like so:

A screenshot of a computer

Description automatically generated

In databricks now click on create metastore:

A screenshot of a computer

Description automatically generated

Put in a name for the metastore and then the path syntax is shown above. The adls gen 2 path is as stated, and the access connector id is the resource id as shown below:

A screenshot of a computer

Description automatically generated

We can now put the details into databricks to create the metastore. So the details are as follows:

A screenshot of a computer

Description automatically generated

After clicking create, we can now see all of the adls containers that are available for us to use. The pricing tier must be premium and the region should be where you are.

A screenshot of a computer

Description automatically generated

Click on the workspace/container that you want and click assign and then enable.

A screenshot of a computer

Description automatically generated

You will then get this message below:

A screenshot of a computer

Description automatically generated

Workspace

In workspace on the left, you can see the workspaces that you have.

A screenshot of a computer

Description automatically generated

By clicking on the workspace, you can configure it.

A screenshot of a computer

Description automatically generated

In the data tab, you can create or delete metastores

A screenshot of a computer

Description automatically generated

You can also manage user access in user management:

A screenshot of a computer

Description automatically generated

Let’s now open the workspace

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Creating a SQL Warehouse

A sql warehouse attaches compute and then performs queries on stored tables and views.

Go to create SQL warehouse

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

DBU represented a combination of CPU, memory, and IO resources. The higher the DBU, the more cost there is. You can set the manual cluster. Scaling clusters leads to more worker nodes when required, but this is more costly. We can choose serverless, pro, or classic warehouse type. Make sure unity catalog is checked.

A screenshot of a computer

Description automatically generated

You can add permissions for the cluster.

A screenshot of a computer

Description automatically generated

Add that all users can use the serverless SQL data warehouse for permissions.

A screenshot of a computer

Description automatically generated

You can monitor the SQL warehouse below.

A screenshot of a computer

Description automatically generated

Also, you can stop the cluster with the stop button below.

A screenshot of a computer

Description automatically generated

Sql Editor

A screenshot of a computer

Description automatically generated

When you run sql queries, you select the SQL warehouse to run the queries against.

Adding a query

You can add a new query to the workspace as below. For example, in the user folder, you can add a new query by clicking add.

A screenshot of a computer

Description automatically generated

In the catalog, you can use the sample data to query in the example below.

A screenshot of a computer

Description automatically generated

If I attach a SQL warehouse by clicking the start button below:

A screenshot of a computer

Description automatically generated

Then I go back to the SQL editor and then attach the serverless SQL warehouse

A screenshot of a computer

Description automatically generated

The sample tables will now all appear:

A screenshot of a computer

Description automatically generated

In databricks the schemas are known as the database, and we can see the schemas in the sample catalog. They contain tables and views. So in thesamples catalog, we have the nyc taxi schema/database, and we have the trips table. We can then expand that and see the columns in the trips table.

A screenshot of a computer

Description automatically generated

We use SQL to query as below. Make sure the right catalog and schema are selected above before querying.

A screenshot of a computer

Description automatically generated

The above shows the top 1000 records, but if we unselect the LIMIT 1000 as below, then we find that we can see all the rows in the table.

A screenshot of a computer

Description automatically generated

We can also unselect ‘run all statements’ if we want to run some statements and not all.

Saving a query

I can save a query by going to the top right and then clicking save. Navigate to the users folder and your name, and then save the query. The query is stored here:

A screenshot of a computer

Description automatically generated

Using schemas and catalogs

We can change schemas and catalogs by using the following commands:

USE Schema

USE catalog.

Scheduling queries

You can schedule queries constantly by using the schedule button.

A screenshot of a computer

Description automatically generated

The query can be scheduled to run every day, hour, second, or any frequency you want.

A screenshot of a computer

Description automatically generated

Under query history, we should see the query that was scheduled if it executes. All other queries can be seen here too.

A screenshot of a computer

Description automatically generated

Make sure you turn off the schedule if you don’t want to incur costs.