**Databricks Operations Run Book**

Revision History

| Date | Version | Description | Author |
| --- | --- | --- | --- |
| 6/30/2023 |  |  | Databricks Admin Team |
| 7/18/2023 | 1.1 | Added below sections:  Configure Databricks CLI  Grant access to Job cluster policies | Databricks Admin Team |
| 8/17/2023 | 1.2 | Service Principals, PAT Tokens and Keyvault Secrets | Databricks Admin Team |
|  |  |  |  |
|  |  |  |  |

**Contents**

[Grant View access to User Groups for the workflows: 2](#_Toc1184834033)

[Use CLI to grant access to multiple workflows: 4](#_Toc204486879)

[Grant Manage Run access to the Ops teams to the workflows: 6](#_Toc886108376)

[Add User email/Group DL to workflow for notifications: 6](#_Toc1247372019)

[Cancel/Kill the running workflow/job: 7](#_Toc1733314828)

[Copy config files from local to ADLS landing zone: 8](#_Toc733466631)

[Create Folder in Storage account: 10](#_Toc1804744260)

[Issue with Different owner for the Table and View 11](#_Toc393666361)

[Run the notebook for specific Databases to change the owner: 12](#_Toc1215713031)

[Change the owner of the tables manually: 12](#_Toc1650219526)

[Request to add AD group to SCIM App: 15](#_Toc795551053)

[Creating Databases: 16](#_Toc2105727984)

[Granting access to the Databases: 17](#_Toc225929446)

[Delete Databases: 21](#_Toc1571059563)

[Delete a workflow: 22](#_Toc380989295)

[List Files in ADLS Folder: 22](#_Toc1916963721)

[BDR JOB: 24](#_Toc714322104)

[Sample Code to do Few Tasks: 26](#_Toc844654094)

[Databricks Cluster Report: 26](#_Toc1417101753)

[Command to list the jobs which are using policy: 27](#_Toc1194536150)

[Information needs to get from new project or team: 28](#_Toc87603805)

[Configure Databricks CLI: 30](#_Toc1607193049)

[Get PAT Token: 31](#_Toc2054372741)

[Grant Access to Job cluster policy: 32](#_Toc2132745403)

[Grant access to Policy: 32](#_Toc1954063445)

[Grant access to the scope: 33](#_Toc2052602099)

[Issues and Solutions: 34](#_Toc1621715859)

[Cluster Issue - Driver is up but is not responsive, likely due to GC: 34](#_Toc2002195196)

[Insufficient privileges: 36](#_Toc551573124)

[Job Failed due to out of memory: 37](#_Toc1764792176)

[Stage Failures: 37](#_Toc379936234)

[Worker node lost: 37](#_Toc1505841873)

[Relative Path Issue: 38](#_Toc1654463058)

[Remote RPC client disassociated: 39](#_Toc784868347)

[Out Of Memory due to collect() and dbutils.fs.put() functions: 40](#_Toc1464441244)

[Job Cluster launch Issue 41](#_Toc1822186318)

[Max out the Task Runs in workspace 43](#_Toc662684661)

[Inactive User Issue: 43](#_Toc1224542779)

[Update node type in Policies not affecting in some existing workflows 44](#_Toc73587364)

[ColumnMappingUnsupportedException: 44](#_Toc585408388)

[Parquet does not support timestamp.See HIVE-6384: 44](#_Toc1834258065)

[Bad Request Error while uploading file and creating table using SQLwarehouse: 45](#_Toc1940236328)

[Service Principals, PAT Tokens and Keyvault Secrets: 46](#_Toc298504670)

[Third-Party Teams details: 46](#_Toc1915162634)

[DBUTILS Commands: 46](#_Toc2025138639)

[Cleanup the files/subfolders under given folder: 47](#_Toc1397856573)

[Storage Accounts and SPNs: 47](#_Toc1305403968)

[Script to delete inactive users: 48](#_Toc1482709527)

[Database access Request instructions: 49](#_Toc1248108191)

[Moving a schema from one ADLS location to Other: 49](#_Toc1793616463)

[Add Service Principal to workspace: 51](#_Toc73563061)

[Generate PAT Token for Service principal: 52](#_Toc1520272355)

[Init Scripts to refer files in Volumes in Unity Catalog volumes: 53](#_Toc81152809)

[Accessing storage account from Cluster/Job policy: 54](#_Toc1148423893)

[Inbound and Outbound Applications: 55](#_Toc2119961469)

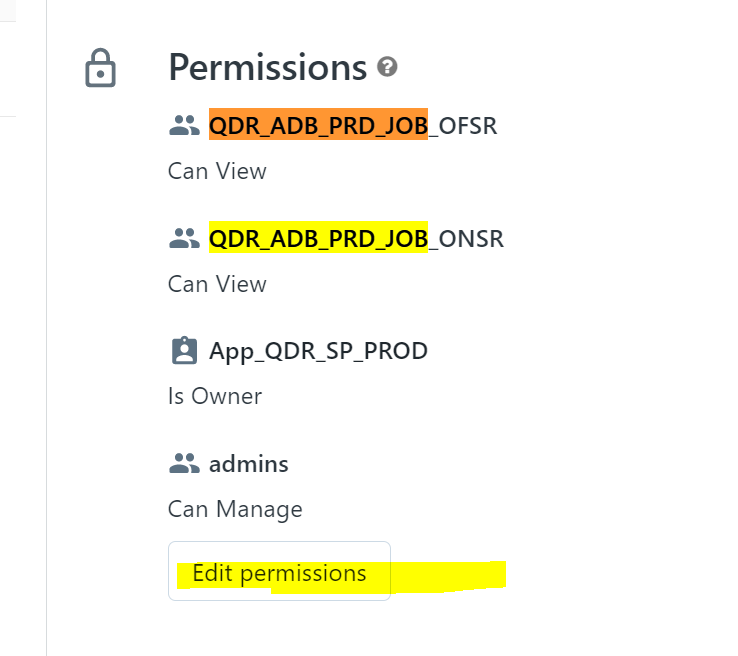
# **Grant View access to User Groups for the workflows:**

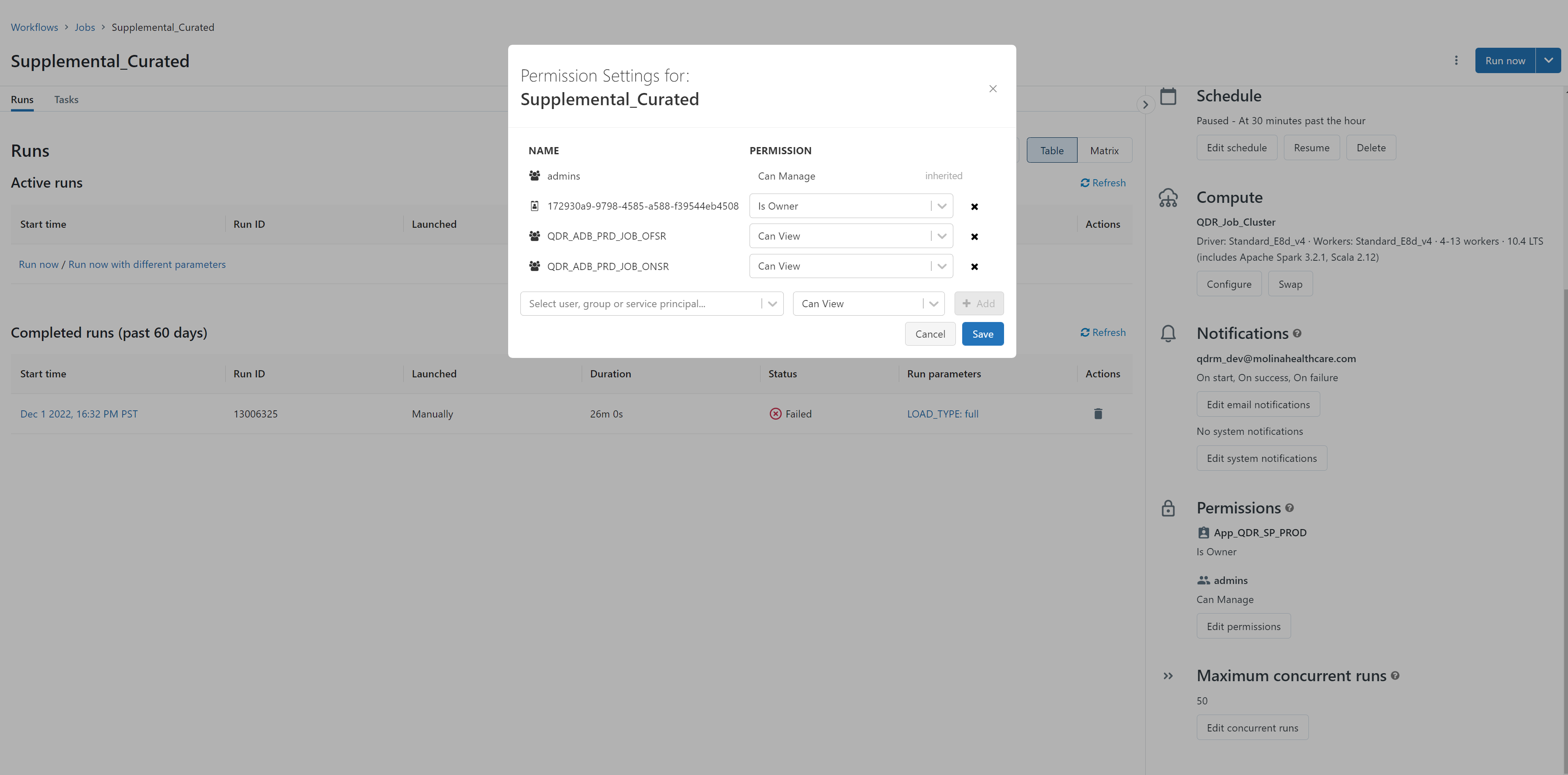
[RITM3408570](https://molina.service-now.com/sc_req_item.do?sysparm_tiny=c22711de1b52a410a72cc80b234bcb9c&sys_id=c8bdee264775a99499004a88036d4351&sysparm_record_row=1)

The above RITM is the sample created by users for view access for logs.

Search the job ID or job name in the workflows.

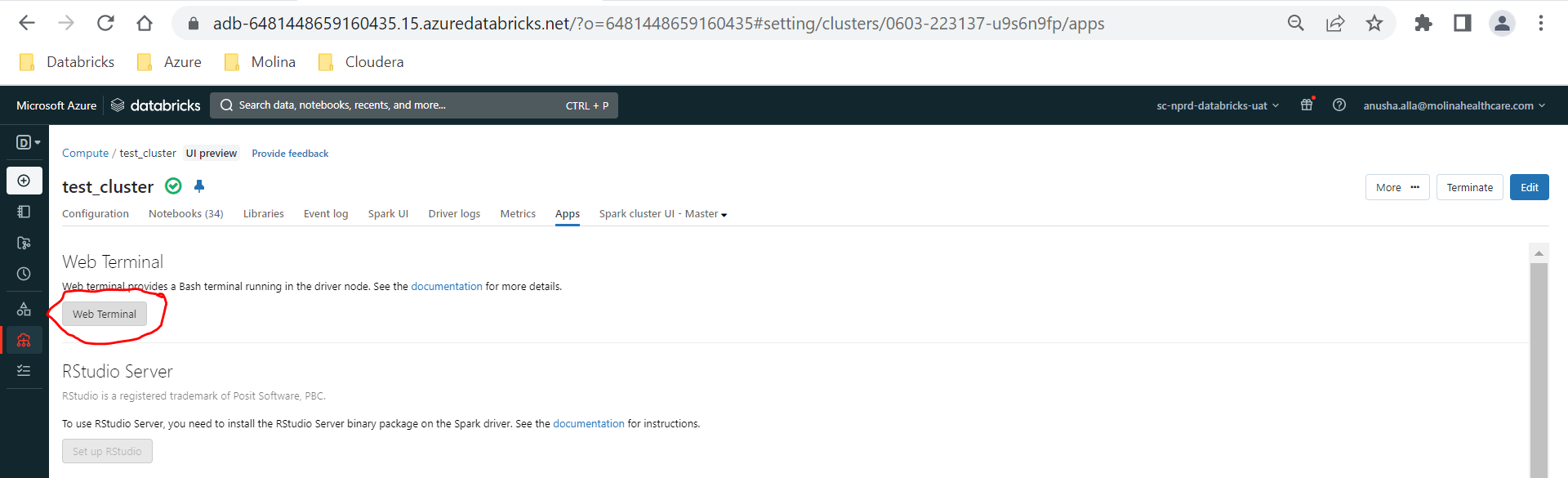
Go will permission in the UI and we can edit permissions and add the ad groups into search filed and give access (view only)





## **Use CLI to grant access to multiple workflows:**

* Login to Databricks workspace
* Start the test\_cluster if it is not up
* Go to apps – click on web terminal



* Make sure databricks-cli is installed. If it is not installed, run the below command to install databricks cli.

pip install databricks-cli

* Run the below commands to export host and personal access token

export DATABRICKS\_HOST=https://adb-3372830949335037.17.azuredatabricks.net

export DATABRICKS\_TOKEN="d\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***"**

To get the token, please click on your name then **User settings**. Under Access Token click **generate new token.** Give the Comment and lifetime. Copy the token that is generated before click done. Once you click done, you cannot retrieve the same token. You would need to regenerate the new token.

* Create a file and put the jobs ids

Sample File:

root@0818-153154-g4hgyu78-10-235-16-129:/databricks/driver# vi jobs\_list

(Enter the list of job ids)

595334013955912

779725684044816

root@0603-223137-u9s6n9fp-10-235-176-58:/databricks/driver# cat jobs\_list

595334013955912

779725684044816

root@0603-223137-u9s6n9fp-10-235-176-58:/databricks/driver#

* Create a json file and put the AD groups and Permissions:

Sample File:

root@0818-153154-g4hgyu78-10-235-16-129:/databricks/driver# vi jobs\_acl.json

{

"access\_control\_list": [

{"group\_name": "ADL\_AzuUAT\_ENT\_OFS\_R",

"permission\_level": "CAN\_VIEW"},

{"group\_name": "ADL\_AzuUAT\_L0UL\_OFSR",

"permission\_level": "CAN\_VIEW"}

]

}

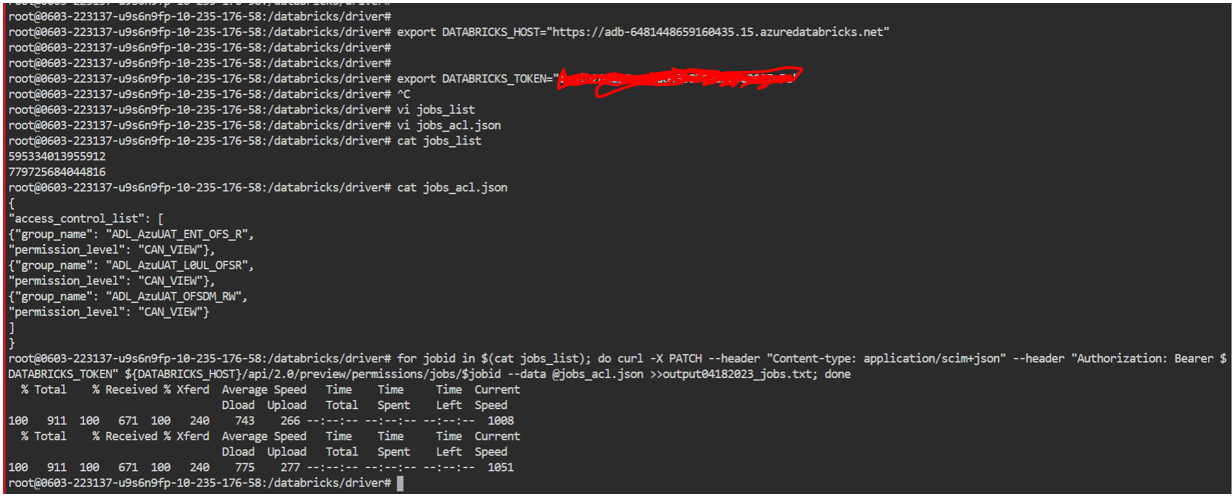
Shift :wq (to save the file)

root@0603-223137-u9s6n9fp-10-235-176-58:/databricks/driver# cat jobs\_acl.json

root@0603-223137-u9s6n9fp-10-235-176-58:/databricks/driver#

* Run the below command to grant permission to the AD groups on the workflows

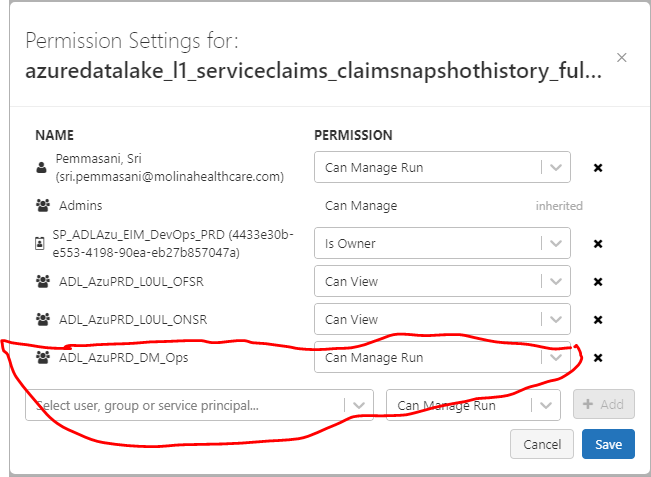
for jobid in $(cat jobs\_list); do curl -X PATCH --header "Content-type: application/scim+json" --header "Authorization: Bearer $DATABRICKS\_TOKEN" ${DATABRICKS\_HOST}/api/2.0/preview/permissions/jobs/$jobid --data @jobs\_acl.json >>output0331\_jobs.txt; done



## **Grant Manage Run access to the Ops teams to the workflows:**

Note that we can grant manage run access to the Ops teams to the workflows with proper approvals.

Reference Ticket: RITM3529899

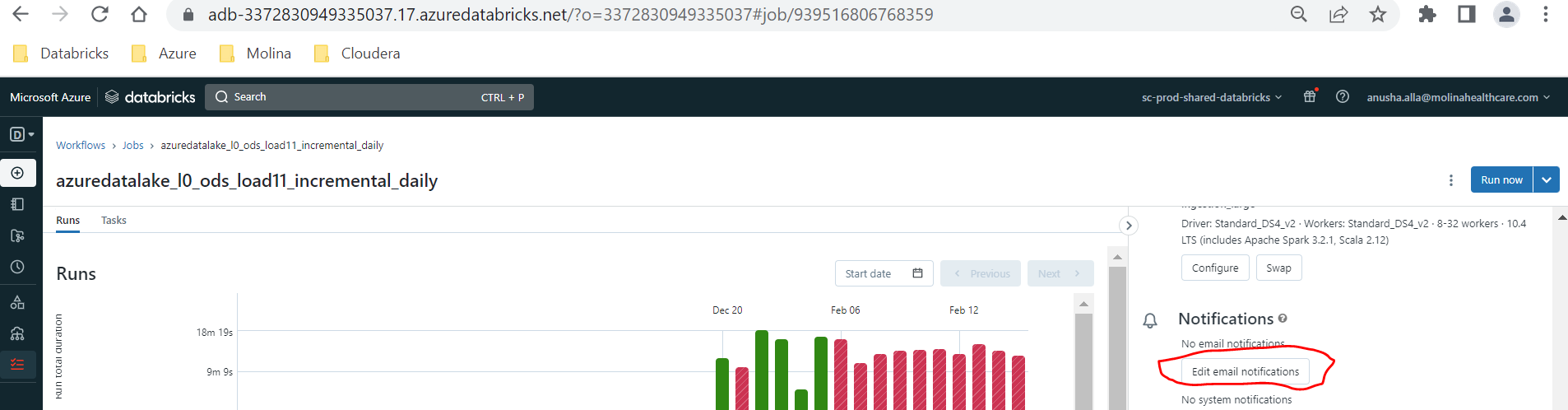


# **Add User email/Group DL to workflow for notifications:**

Reference ticket - RITM3294670

Go to Workflows – Then search for the Job

Click on Edit email Notifications



Add email id and check the success and failure boxes then click confirm.



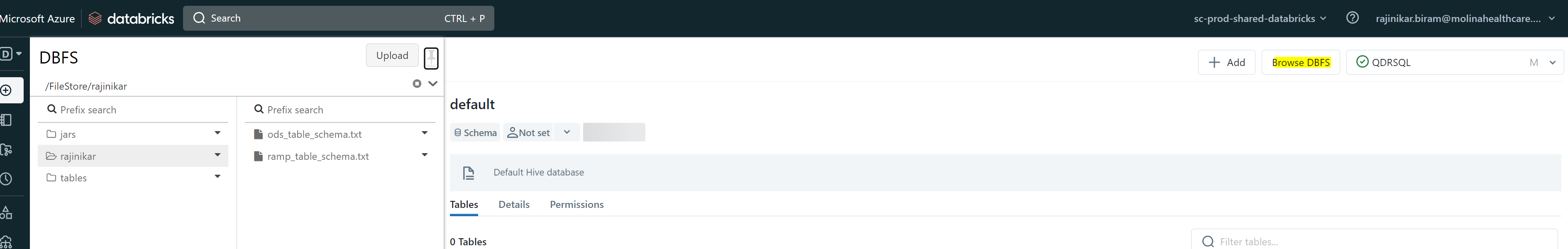
# **Cancel/Kill the running workflow/job:**

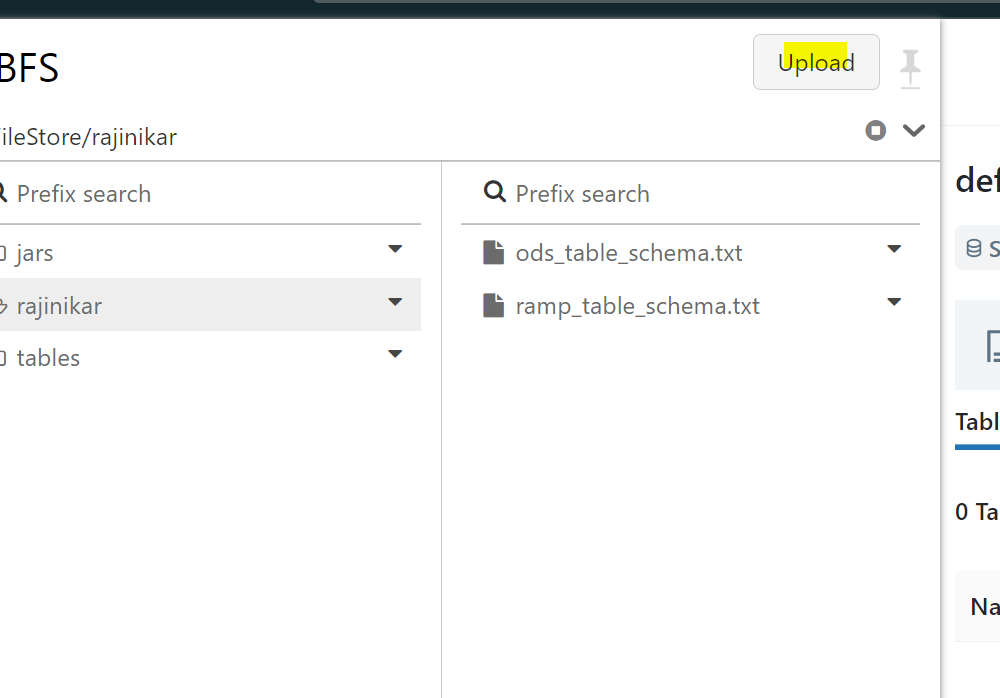
# **Copy config files from local to ADLS landing zone:**

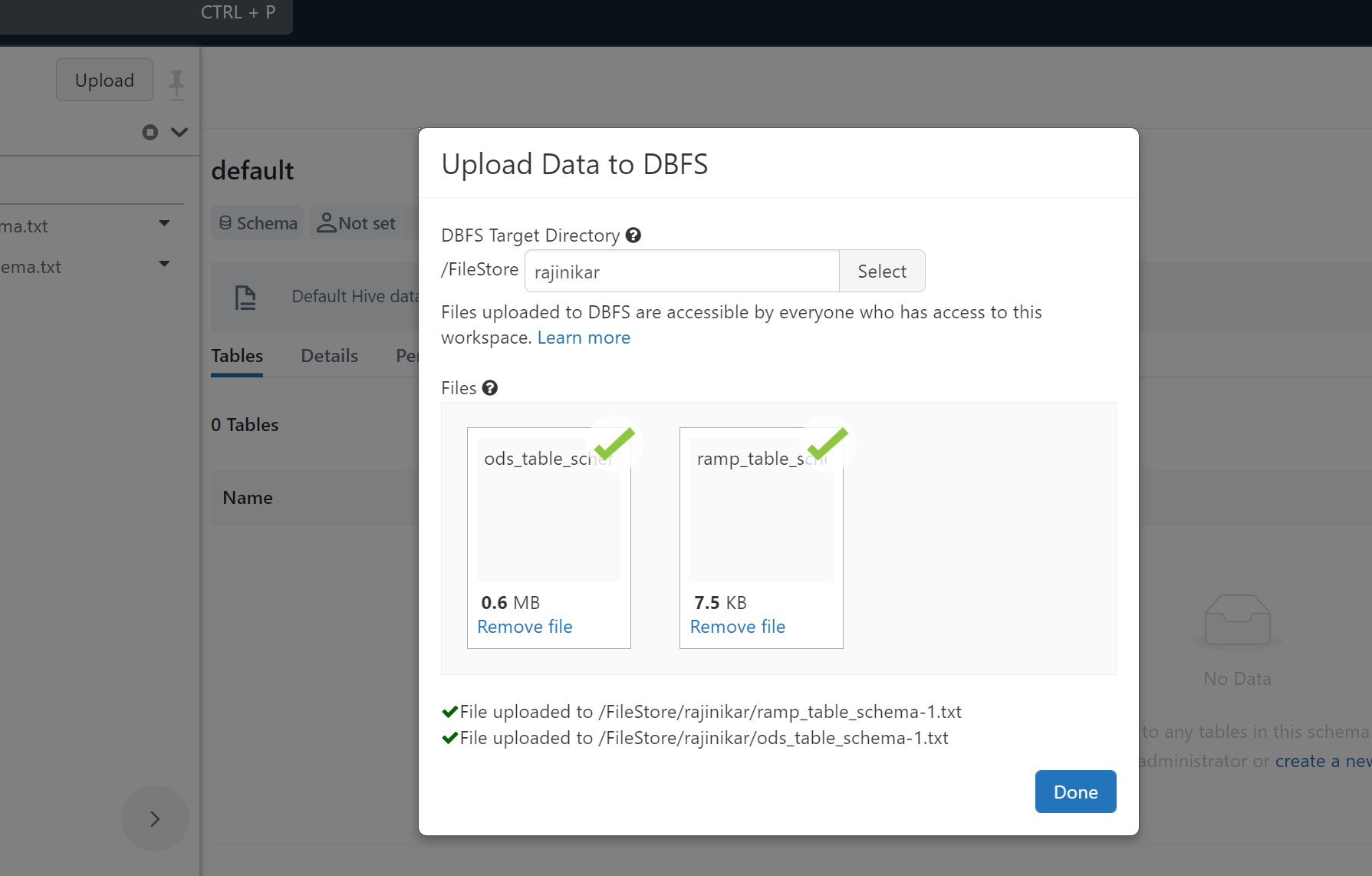
The below steps are used for copying files from local into adls config folders.

User will provide the config files. We need to download into local and upload into dbfs.

Select data in databricks clusters and upload into dbfs path for databricks.







**Sample Commands:**

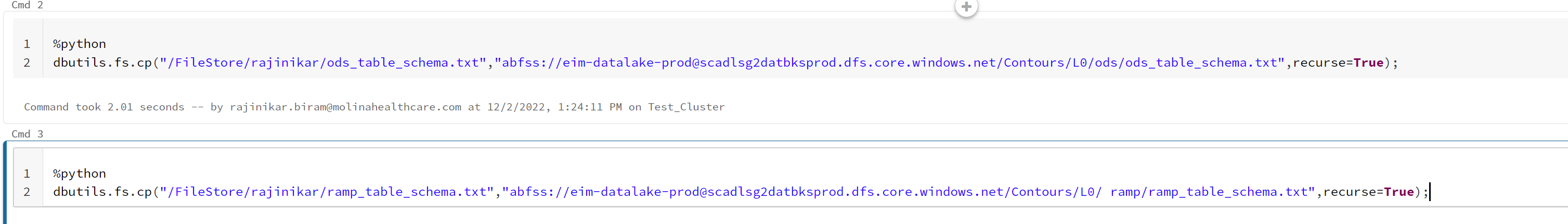
%python

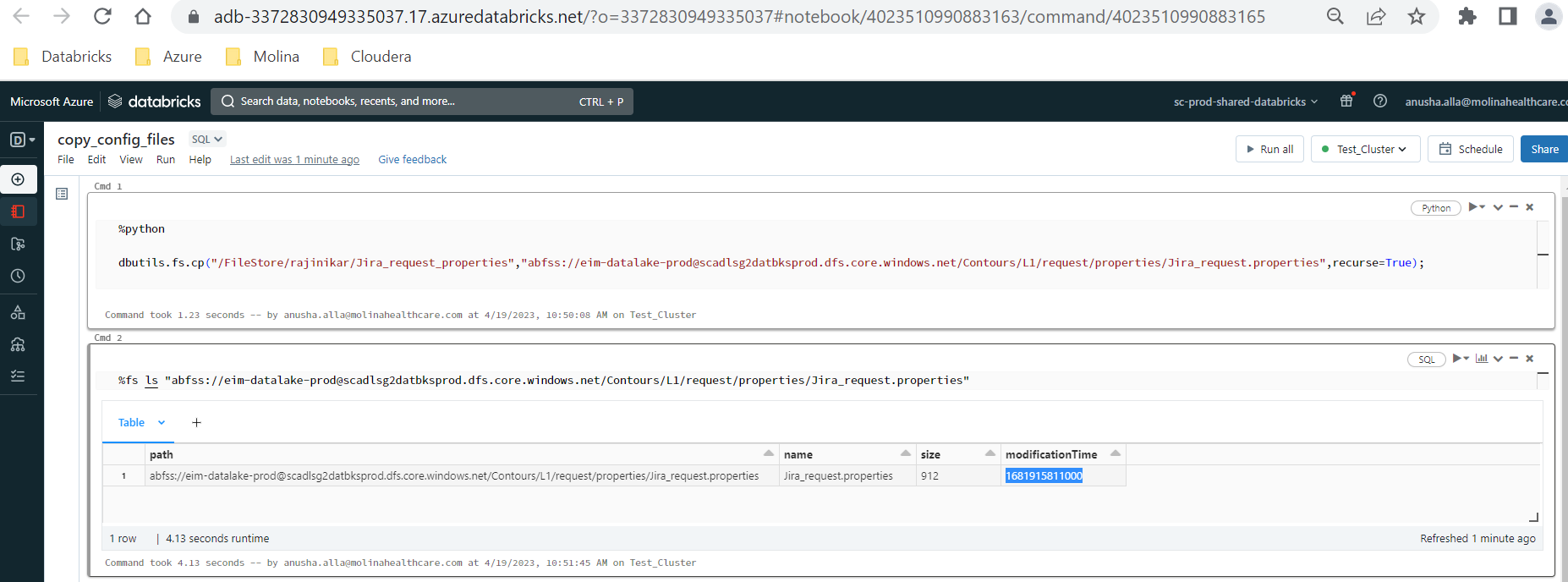
dbutils.fs.cp("/FileStore/rajinikar/ods\_table\_schema.txt","abfss://eim-datalake-prod@scadlsg2datbksprod.dfs.core.windows.net/Contours/L0/ods/ods\_table\_schema.txt",recurse=True);

%python

dbutils.fs.cp("/FileStore/rajinikar/ramp\_table\_schema.txt","abfss://eim-datalake-prod@scadlsg2datbksprod.dfs.core.windows.net/Contours/L0/ ramp/ramp\_table\_schema.txt",recurse=True);

dbutils.fs.cp("/FileStore/rajinikar/**Jira\_request\_properties**","abfss://eim-datalake-prod@scadlsg2datbksprod.dfs.core.windows.net/Contours/L1/request/properties/**Jira\_request.properties**",recurse=True);





**Create Container in Storage account:**

Raise a request with azure ops below is the Link for container creation:

https://molina.service-now.com/com.glideapp.servicecatalog\_cat\_item\_view.do?v=1&sysparm\_id=97ca1240db6a3700c017f482ba9619b2&sysparm\_link\_parent=6084261edb627740c017f482ba9619e9&sysparm\_catalog=e0d08b13c3330100c8b837659bba8fb4&sysparm\_catalog\_view=catalog\_default

Go to the link and edit the fields as per below sample ritm.

Sample RITM: RITM3940959

Azure Resource Group Name:

|  |  |
| --- | --- |
| DEV | SHAREDSERVICE-DATABRICKS-DEV |
| UAT | SHAREDSERVICE-DATABRICKS-UAT |
| PROD | SHAREDSERVICE-DATABRICKS-PROD |

# **Create Folder in Storage account:**

Go to workspace > create notebook

To check the existing folders:

dbutils.fs.ls("abfss://outbound-zone@scadlsg2datbksobzoneprod.dfs.core.windows.net/HiLabs/files/inbound/")

To create new folder (salesforce)

dbutils.fs.mkdirs('abfss://outbound-zone@scadlsg2datbksobzoneprod.dfs.core.windows.net/HiLabs/files/inbound/salesforce/')

**Issue with Different owner for the Table and View:**

Error 1:

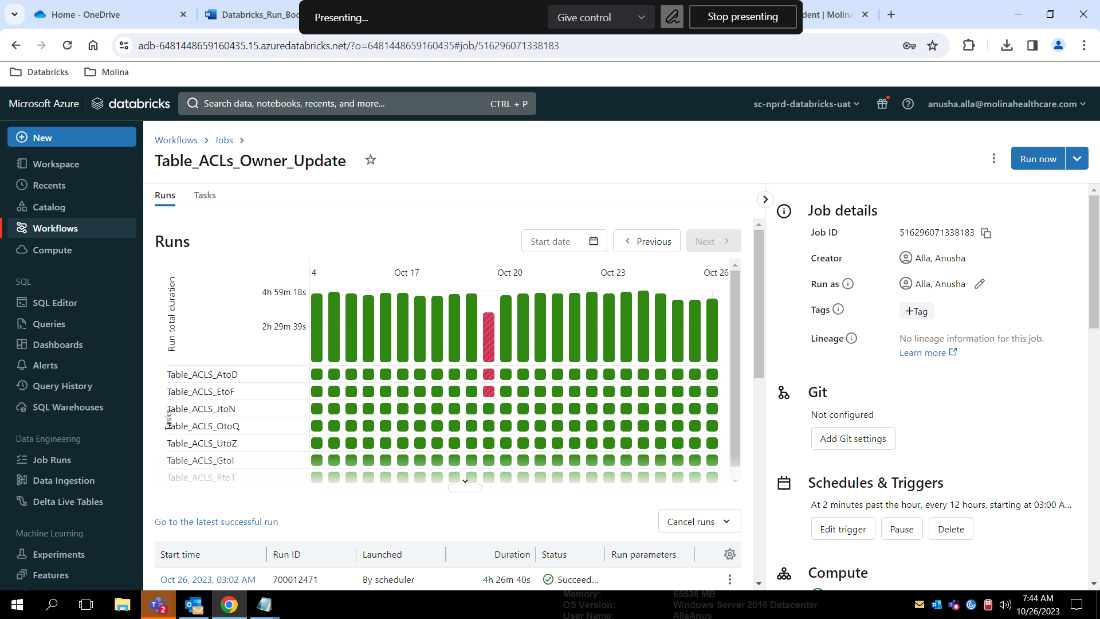
User cannot SELECT on table `enterprise\_db\_secure`.`membermonth` because: --The owner of table `enterprise\_db`.`membermonth` is different from the owner of table `enterprise\_db\_secure`.`membermonth`.

Error2:

Insufficient privileges due to Underlying permissions

**IN UAT: Run Table\_ACLs to change the owner of the tables and views. This job is already scheduled to run 2 times per day. Please note if job fails, please check with user which tables they are facing issue with and change the owner manually to admins. Steps are given below.**

Table\_ACLs\_Owner\_Update:



**Note:** Whenever we create schema/db please change the owner of the schema to AD group [RG\_NIX\_Hadoop\_Admins\_Cust](https://adb-6481448659160435.15.azuredatabricks.net/?o=6481448659160435#setting/accounts/groups/724668894660898)

In Prod, Table are owned by Service principal (SPN) so when changing the owner of the table, make sure to choose right SPN.

In Prod, reporting schema/db/tables owner should be AD group. To change the owner of the reporting dbs or tables, please check with Data governance team to find the right AD group and change to it.

## **Run the notebook for specific Databases to change the owner:**

Run the Notebook - /workspace/Admin\_Stuff/Table\_Acls\_against\_SpecificDB

**Note:** Please make sure to use Schema\_ACLs\_Cluster as Table ACLs are enabled only on this cluster with high concurrency cluster mode.

In command/cell 3, give the database names in db\_list as given below. Below databases were given as example. Please give the database name where you would need to change the owner of the tables and views.

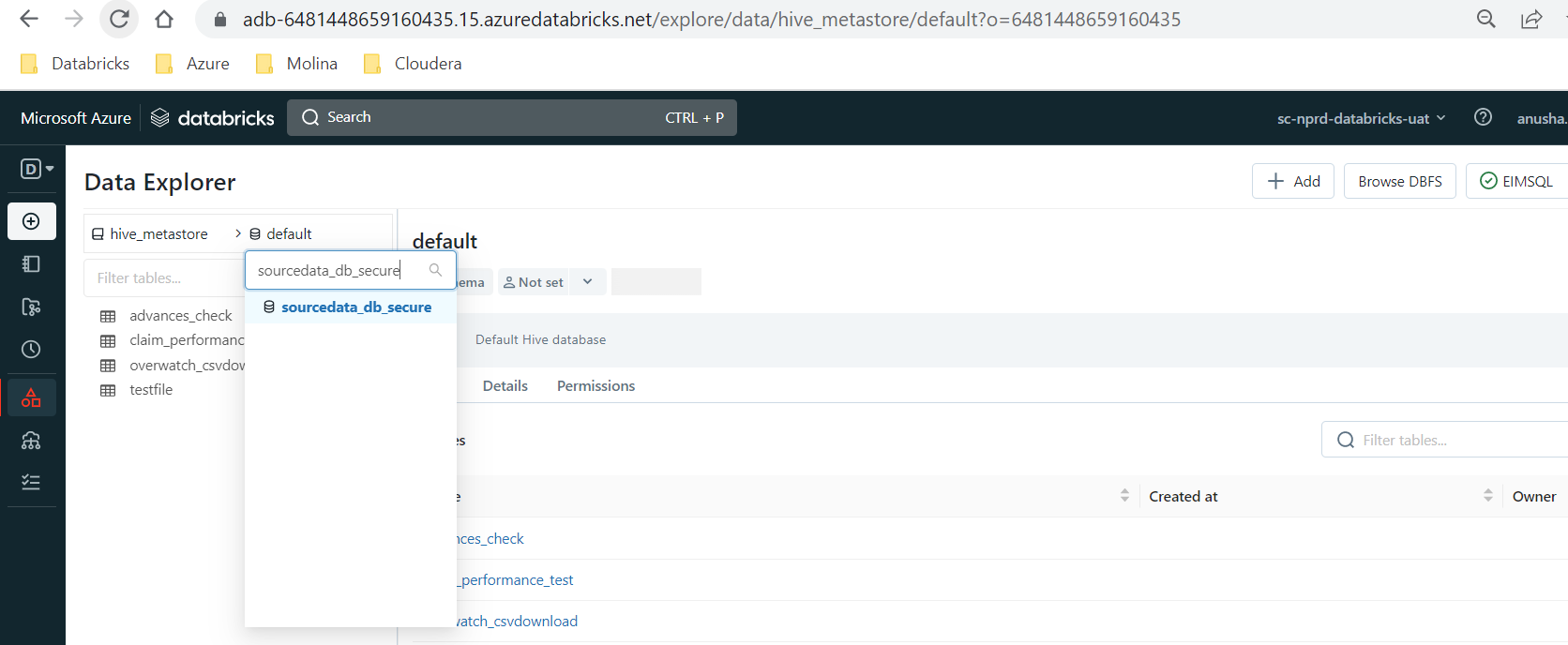
**Example:** For one database- db\_list = ['appeals\_db\_secure']

If it is more than one database then give database names with comma separated - db\_list = ['appeals\_db\_secure', 'enterprise\_db\_secure']

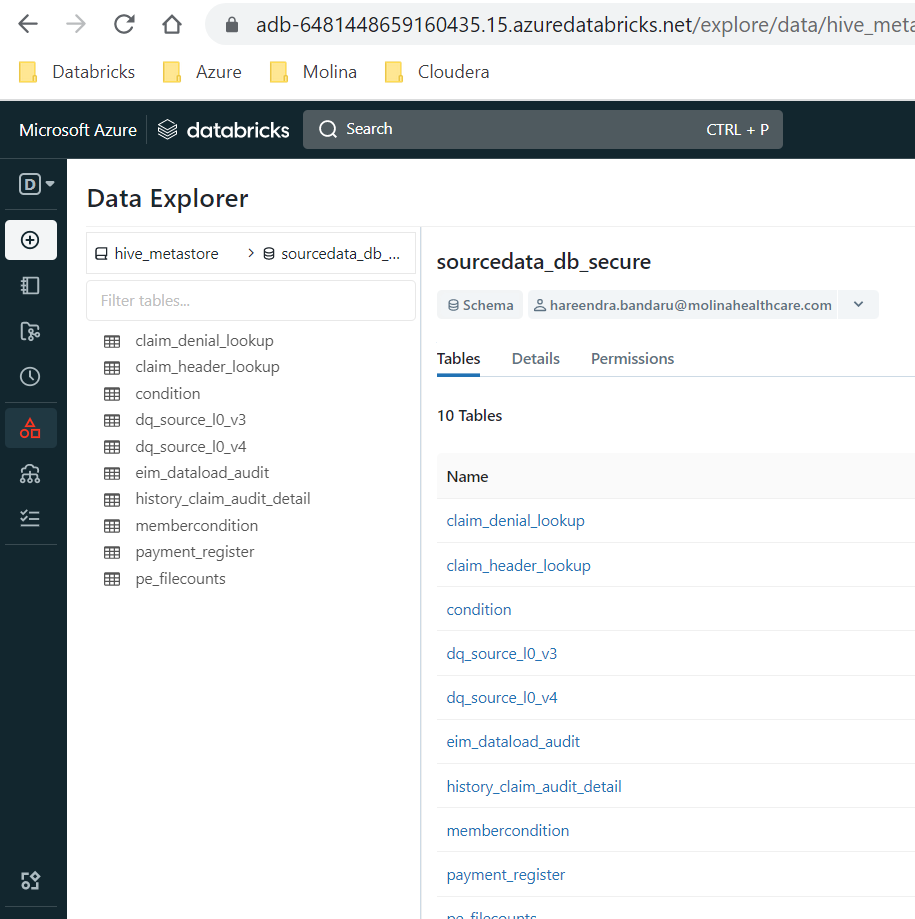
Click Run All.

## **Change the owner of the tables manually:**

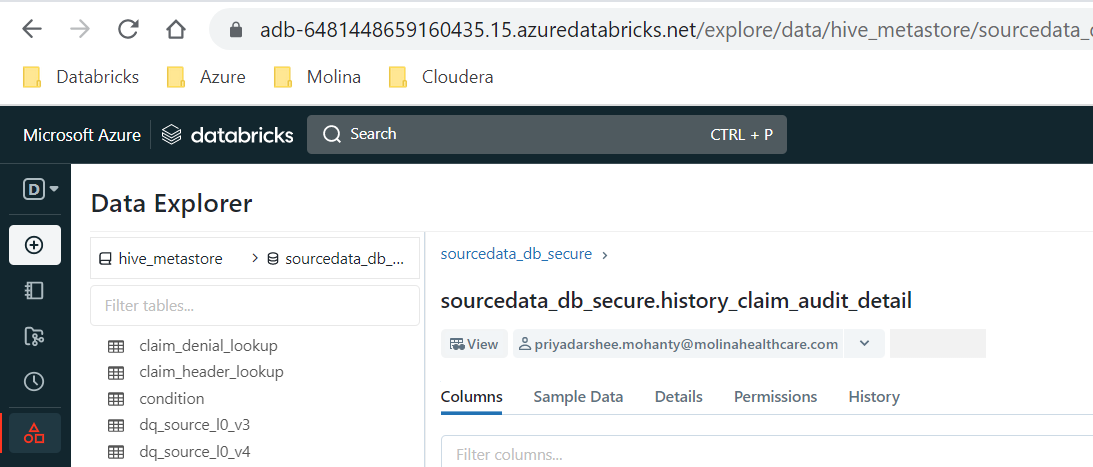
Search for the DB in Data Explorer



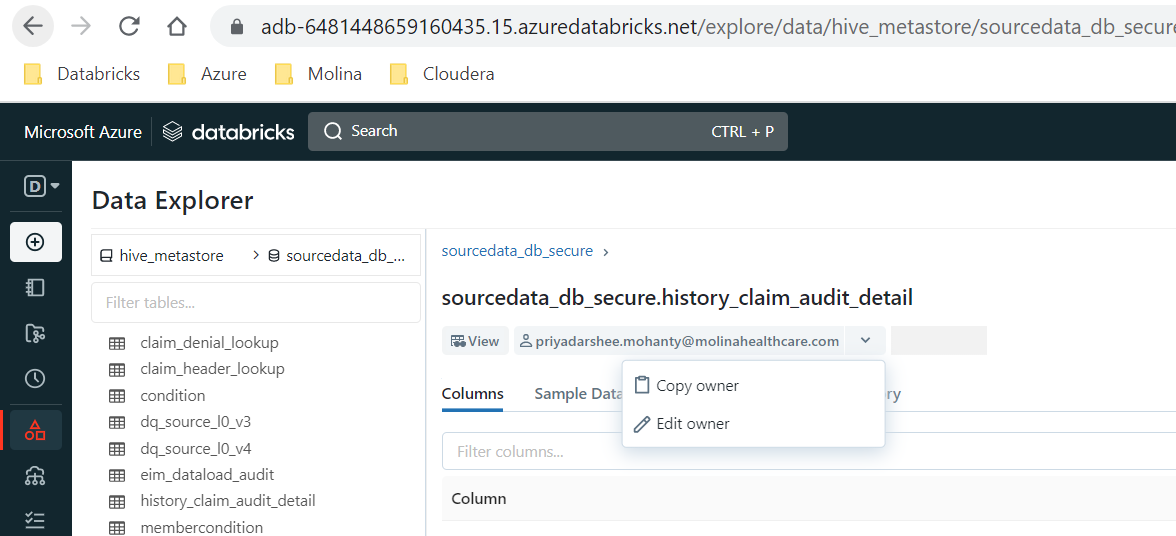
Select DB



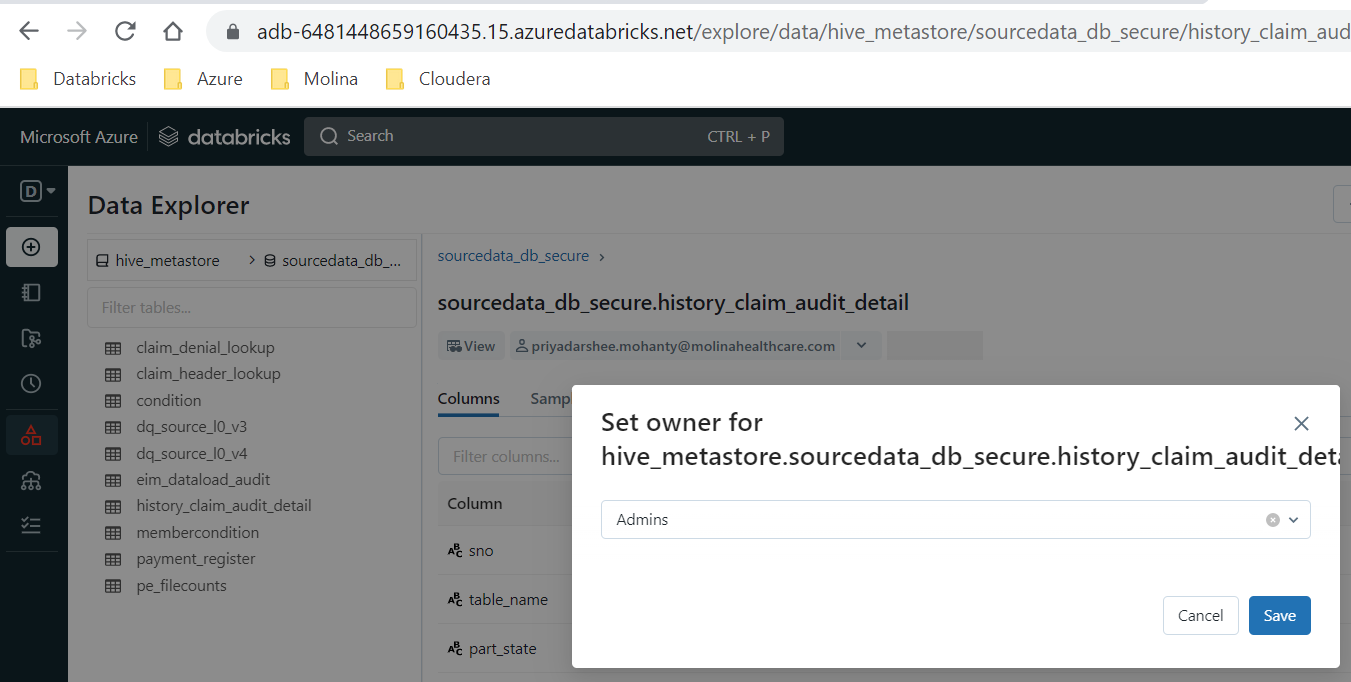
Select the table which is having issue



Click the drop down box for the owner and click edit owner



Search for the admins and click save



Make sure you see admins as the owner now.



# **Request to add AD group to SCIM App:**

If there is any new groups, we would need to open a request to AD APP registration team to add the AD groups to specific environment Databricks SCIM apps.

Please refer RITM3142186

[Azure AD App Registration | Molina Service Management Platformyreaeeaewswc (service-now.com)](https://molina.service-now.com/com.glideapp.servicecatalog_cat_item_view.do?v=1&sysparm_id=5d0d310edb0e0c90e787a026ca9619bf)

|  |  |
| --- | --- |
| Environment | SCIM Connector |
| DEV | SC-NPRD-Databricks-DEV |
| UAT | SC-NPRD-Databricks-UAT |
| PROD | SC-PROD-Databricks |

# **Creating Databases:**

Refer the request – RITM3192199

Please note that all L0 and L1 schemas should be in Normal Storage account and L2 and Reporting schemas should be in Reporting Storage account.

**DB Type and ADLS Location:**

|  |  |  |
| --- | --- | --- |
| **Application** | **Environment** | **L0 and L1 Schemas** |
| EIM | PROD | abfss://eim-datalake-prod@scadlsg2datbksprod.dfs.core.windows.net/datalake/db |
| UAT | abfss://eim-datalake-uat@scadlsg2datbksuat.dfs.core.windows.net/datalake/db |
| DEV | abfss://eim-datalake-dev@scadlsg2datbksdev.dfs.core.windows.net/datalake/db |
| QDR | PROD | abfss://qdr-qdr-prod@scadlsg2datbksprod.dfs.core.windows.net/ |
| UAT | abfss://qdr-qdr-uat@scadlsg2datbksuat.dfs.core.windows.net/ |
| DEV | abfss://qdr-qdr-dev@scadlsg2datbksdev.dfs.core.windows.net/ |
| QDR RAMP | Prod | abfss://qdr-qdr-prod@scadlsg2datbksprod.dfs.core.windows.net/ramp/ |
| UAT | abfss://qdr-qdr-uat@scadlsg2datbksuat.dfs.core.windows.net/ramp |
| DEV | abfss://qdr-qdr-dev@scadlsg2datbksdev.dfs.core.windows.net/ramp/ |
| QDR MRF | Prod | abfss://qdr-qdr-prod@scadlsg2datbksprod.dfs.core.windows.net/mrf/ |
| UAT | abfss://qdr-qdr-uat@scadlsg2datbksuat.dfs.core.windows.net/mrf/ |
| DEV | abfss://qdr-qdr-dev@scadlsg2datbksdev.dfs.core.windows.net/mrf/ |
| QDR HIE | Prod | abfss://qdr-qdr-prod@scadlsg2datbksprod.dfs.core.windows.net/hie/ |
| UAT | abfss://qdr-qdr-uat@scadlsg2datbksuat.dfs.core.windows.net/hie/ |
| DEV | abfss://qdr-qdr-dev@scadlsg2datbksdev.dfs.core.windows.net/hie/ |

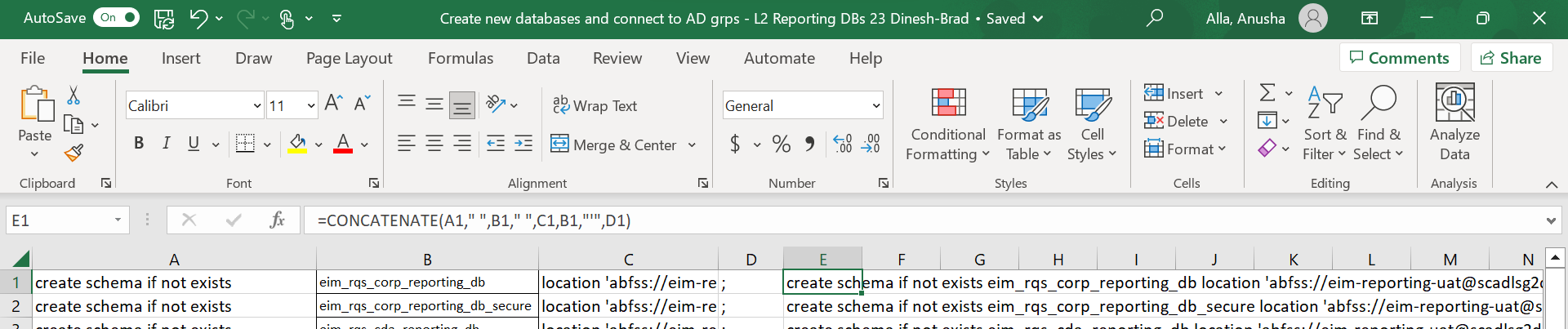
|  |  |  |
| --- | --- | --- |
| **Application** | **Environment** | **L2 and Reporting SChemas** |
| EIM | PROD | abfss://eim-reporting-prod@scadlsg2datbksrptprod.dfs.core.windows.net/datalake/db |
| EIM | UAT | abfss://eim-reporting-uat@scadlsg2datbksrptuat.dfs.core.windows.net/datalake/db/ |

**Query:** Please note to update actual database name, adls storage account and container name in the below query. **Please Note that database location might change based on application as per given above. Initially please check with Rajinikar or Hareendra**.

create schema if not exists <database\_name> location 'abfss://<container>@<storage account>.dfs.core.windows.net/datalake/db/<database\_name>';

**Example:**

create schema if not exists claimshark\_db location 'abfss://eim-datalake-uat@scadlsg2datbksuat.dfs.core.windows.net/datalake/db/claimshark\_db';



E1=CONCATENATE(A1," ",B1," ",C1,B1,"'",D1)

# **Granting access to the Databases:**

We can grant access using SQL Editor or Data explorer

Refer the ticket RITM3203822

Permissions:

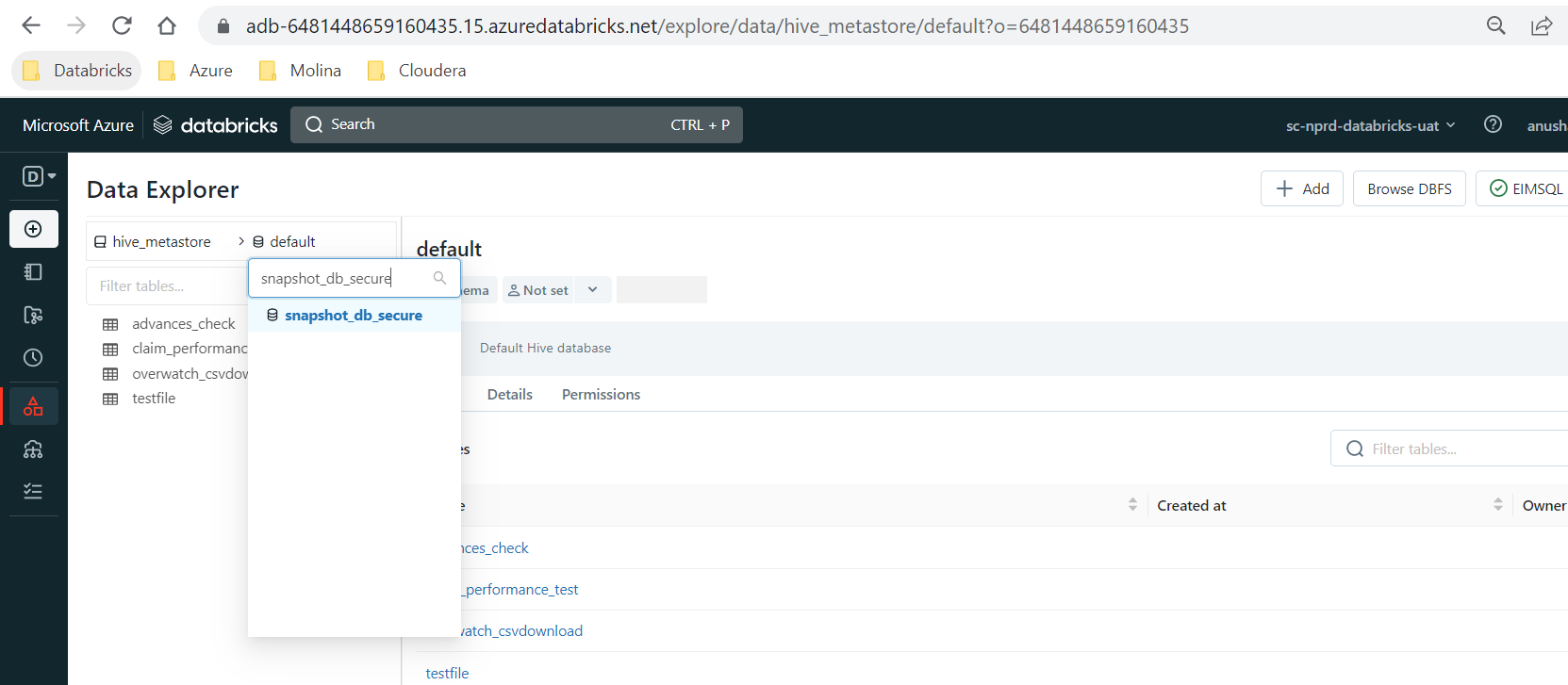
|  |  |
| --- | --- |
| Read-Only Permissions | Write (ALL) Permissions |
| Select, Read\_metadata and usage | Select, Read\_metadata, usage, create and modify |

AD Groups:

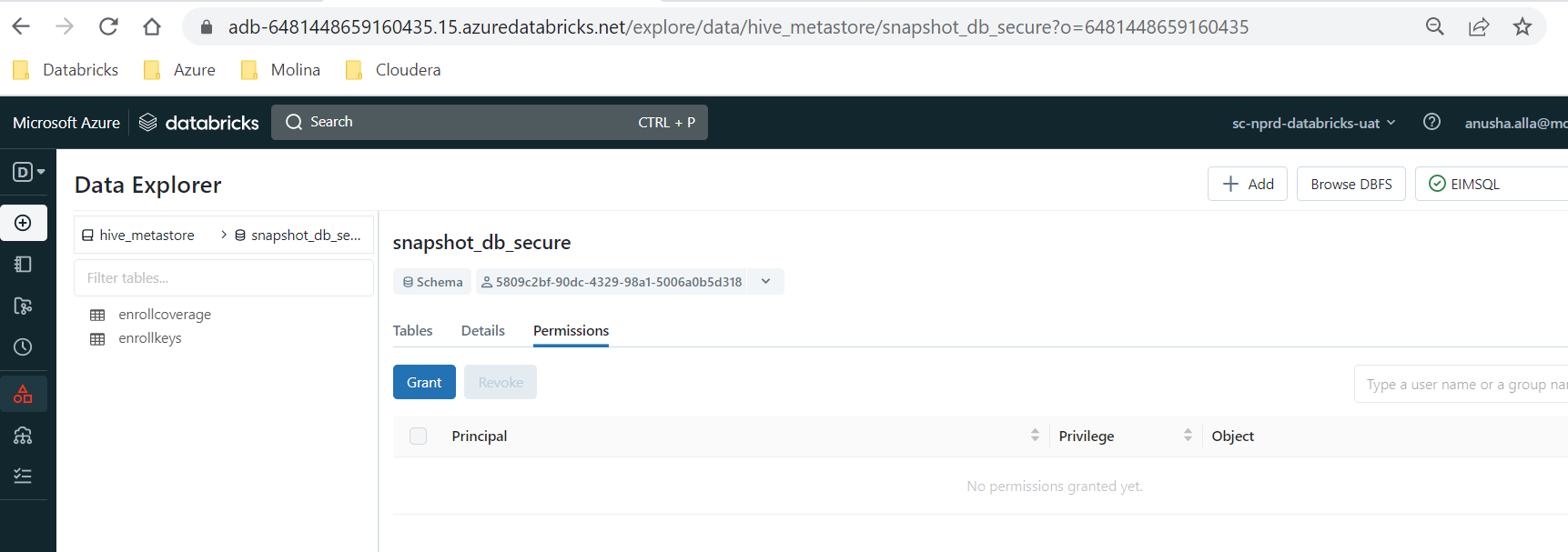
|  |  |
| --- | --- |
| ONS AD Group (Onsite) | OFS AD Group (Offshore) |
| Normal DBs (Example: UMK2\_db) | \_Secure DBs (Example: UMK2\_db\_secure) |

1. Go To Data Explorer – search for Database

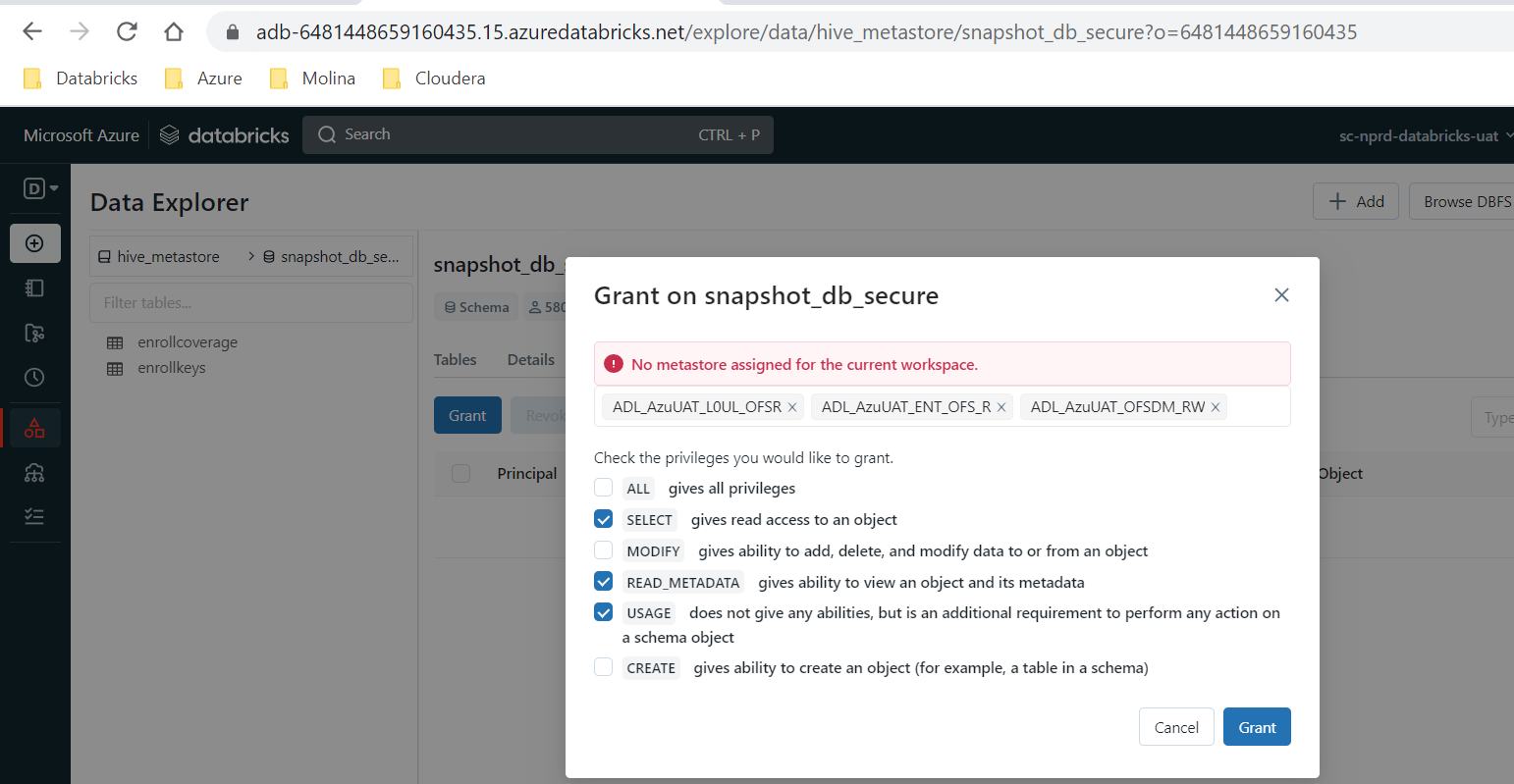
(Please make sure hive\_metastore catalog is selected and one of the SQL warehouses is started and connected).



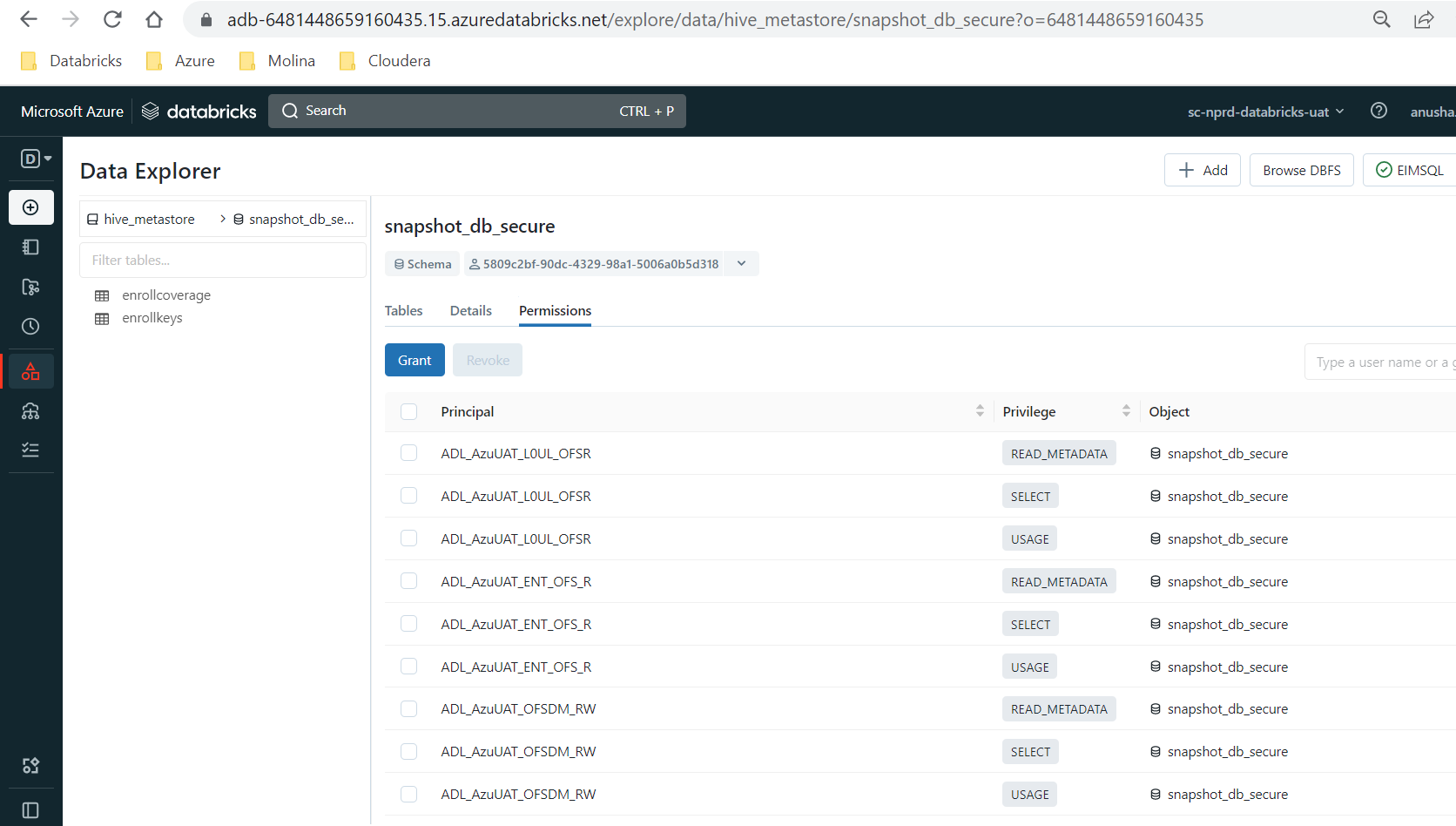
1. Select the database and go to permissions



1. Click on **Grant** and search for ad groups and click on them to add. Then select the permissions bases on read only or read write as given above. Then click **Grant**.



1. Make sure you see groups are added and permissions are given



We can also grant permissions using query in SQL Editior.

**Queries:** Please note to give the actual database name and AD group name in the queris which are in <>.

For Read only:

grant usage,select,read\_metadata on schema <database\_name> to <AD\_Group\_Name>;

For read and write access:

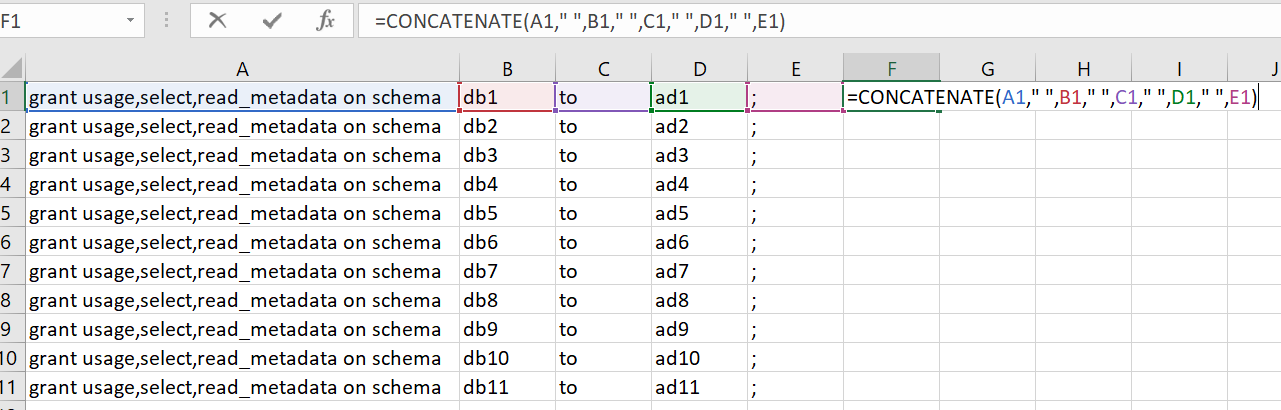
grant usage,select,read\_metadata,create,modify on schema <database\_name> to <AD\_Group\_Name>;

**Examples:**

grant usage,select,read\_metadata on schema claimshark\_db to ADL\_AzuUAT\_L0UL\_ONSR;

grant usage,create,modify,select,read\_metadata on schema claimshark\_db to ADL\_AzuDEV\_L0U\_ONSRW;

When we have more dbs, we can try below to write queries.



# **Delete Databases:**

drop schema if exists dbname;

**Reference Tickets:**

RITM3133694

RITM3175078

For example, delete ipro\_db in UAT.

1. Take the location of Database - abfss://eim-datalake-uat@scadlsg2datbksuat.dfs.core.windows.net/datalake/db/ipro\_db
2. Delete the database using below sql statement

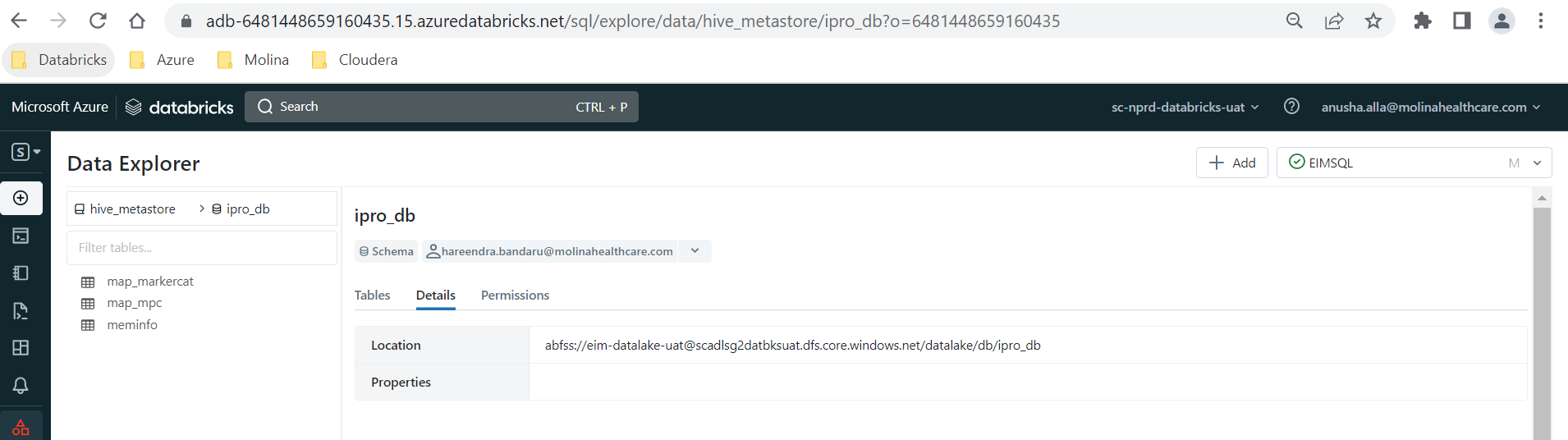
drop schema if exists ipro\_db;

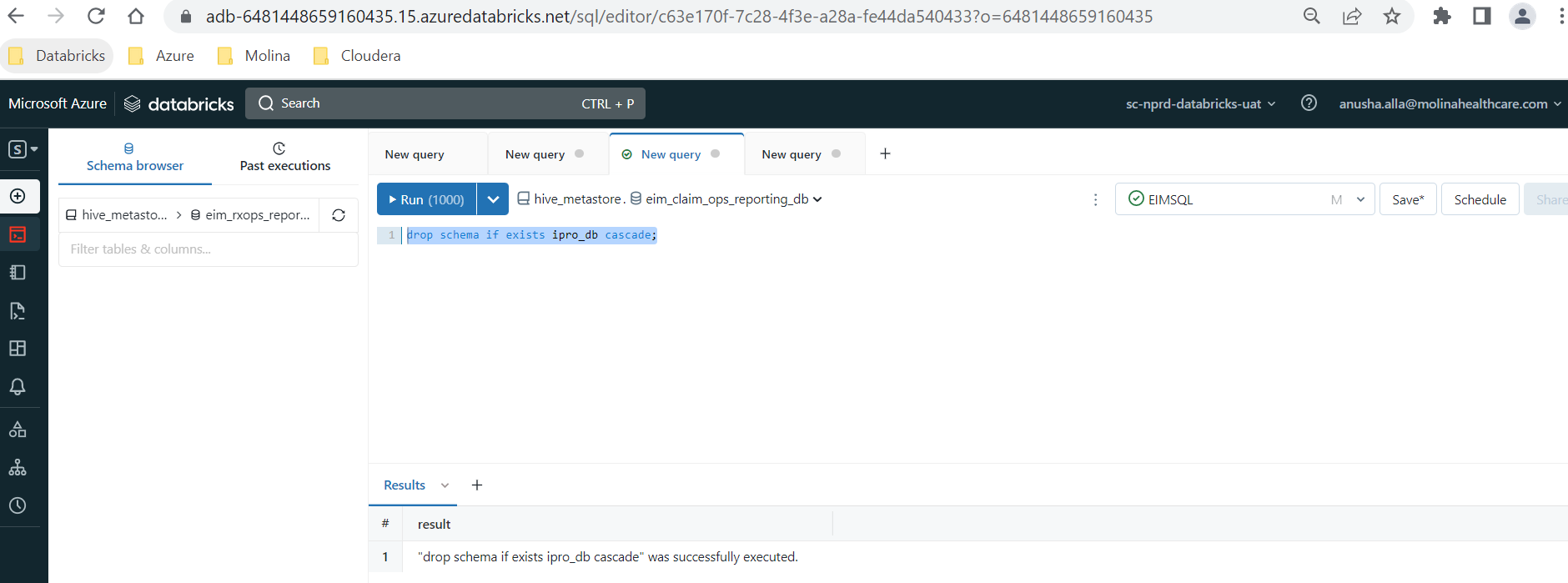
(or)

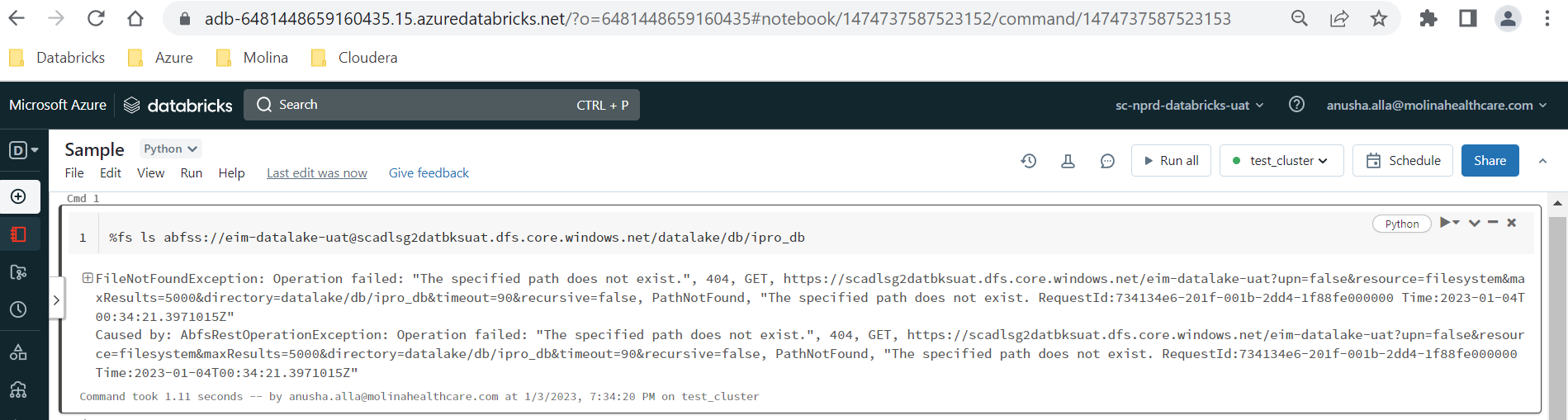
drop schema if exists ipro\_db cascade; --- Please do NOT use cascade option if the tables are not there.

1. Make sure database doesn’t exist in Data Explorer. Also you can make path doesn’t exist by running below in the notebook

%fs ls abfss://eim-datalake-uat@scadlsg2datbksuat.dfs.core.windows.net/datalake/db/ipro\_db

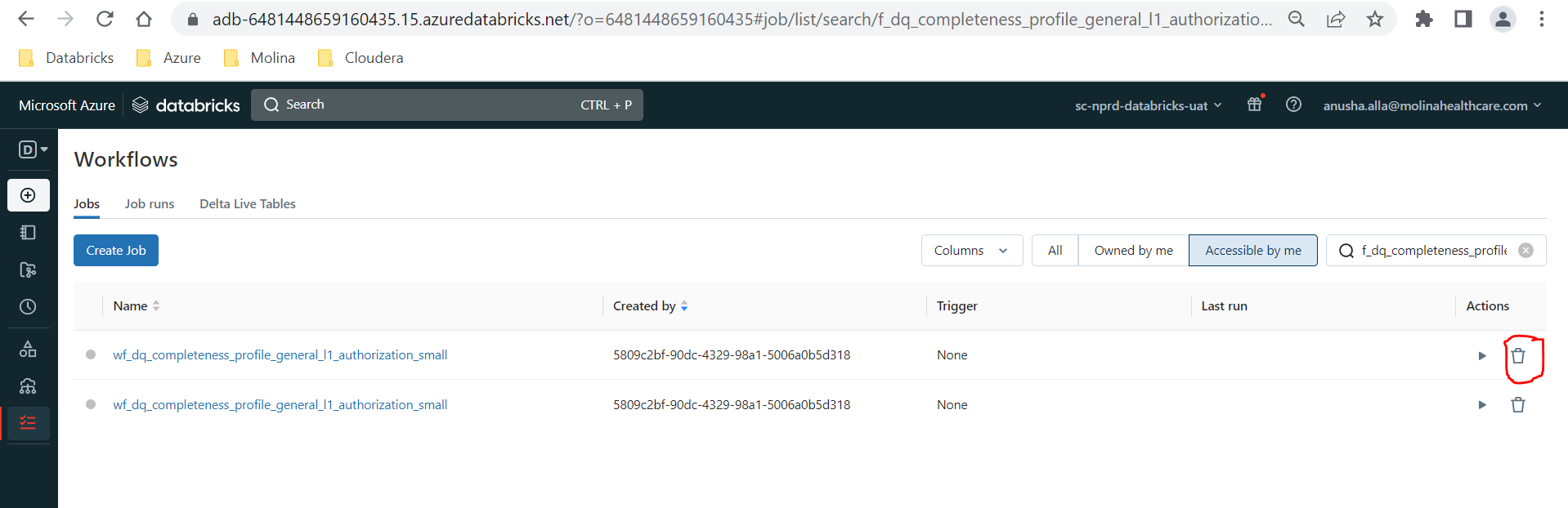






# **Delete a workflow:**

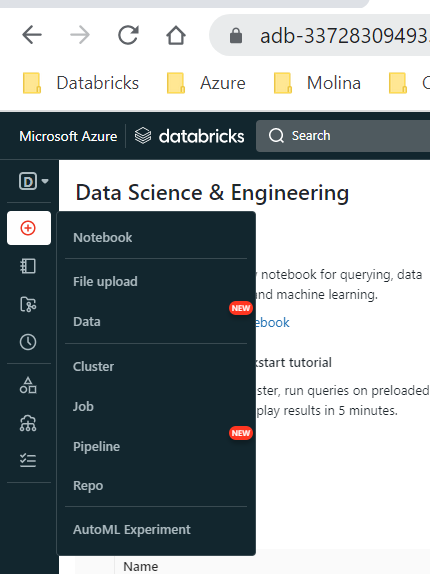
Go to workflows – search for the workflow name or id. Then delete the workflow

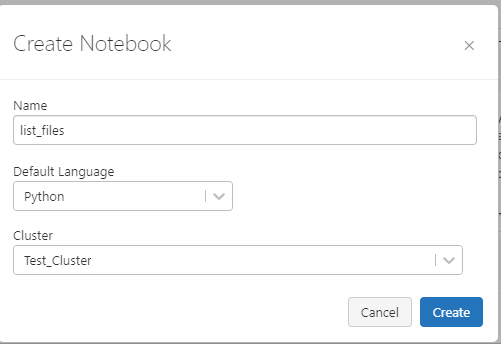


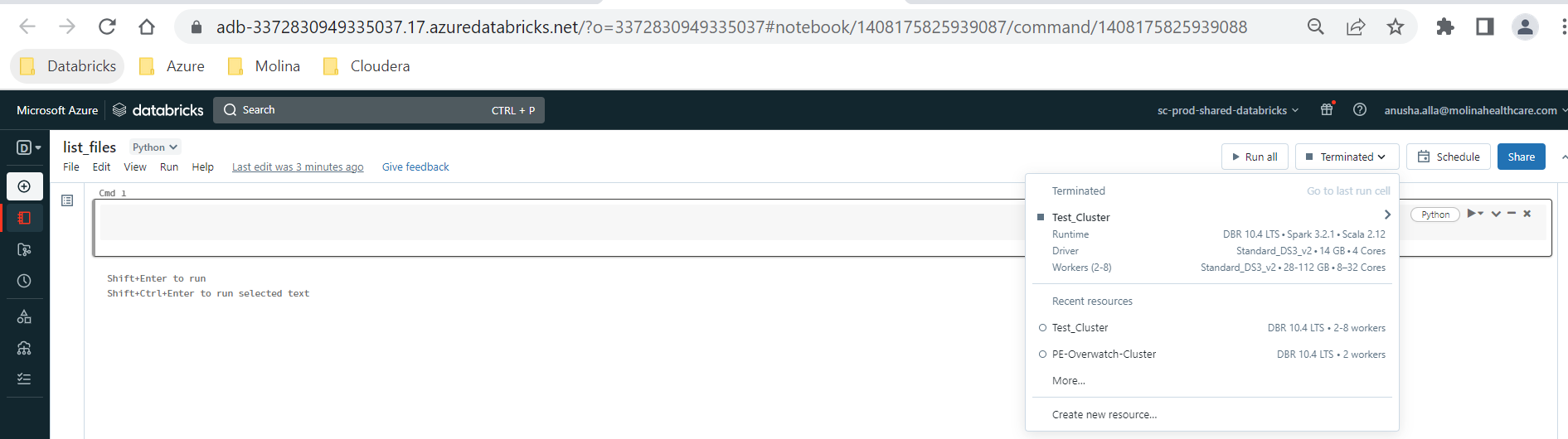
# **List Files in ADLS Folder:**

Create a notebook and attach it to Test\_Cluster. Make sure Test\_cluster is started. If not please go to compute and start the cluster.

You can also start cluster from notebook as given in the screenshots below.







# **BDR JOB:**

* BDR should do always from Cloudera Prod to Azure Prod irrespective of request whether it is for Azure Prod or UAT.
* Once tables are copied to Azure Prod those tables will be automatically synched to Azure UAT by ingestion process
* Shouldn’t do BDR to Azure DEV or UAT.
* Target path will be always Azure Landing zone
* BDR Job name should be unique. Always include ticket number in the BDR job name.

Source: Hive (EDP\_CDH\_PROD)

Destination: dataingestionprdadlsg2

Cloud Root path (Prod Landing zone path): abfss://landingzone@dataingestionprdadlsg2.dfs.core.windows.net/cdh/stagingzone/<db>\_cloudera/

RunasUser: BDR\_PRDCDHADLSSVC01

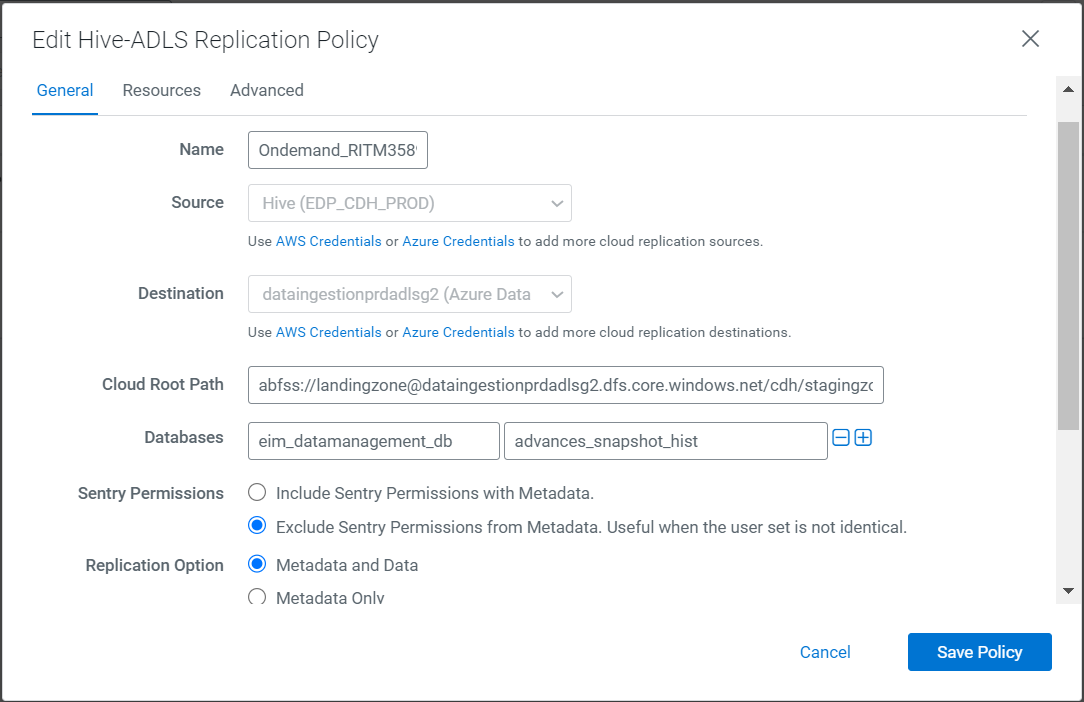
* Select the Max Map Slots and Max Bandwidth based on number of tables and size of the tables. Please make sure you do not overload the server with BDRs and effects performance
* Please always select “delete permanently” in advanced tab which takes care of deleting additional files in azure
* For copying whole DB, you can give the DB name with option shown in below screenshot instead giving all tables

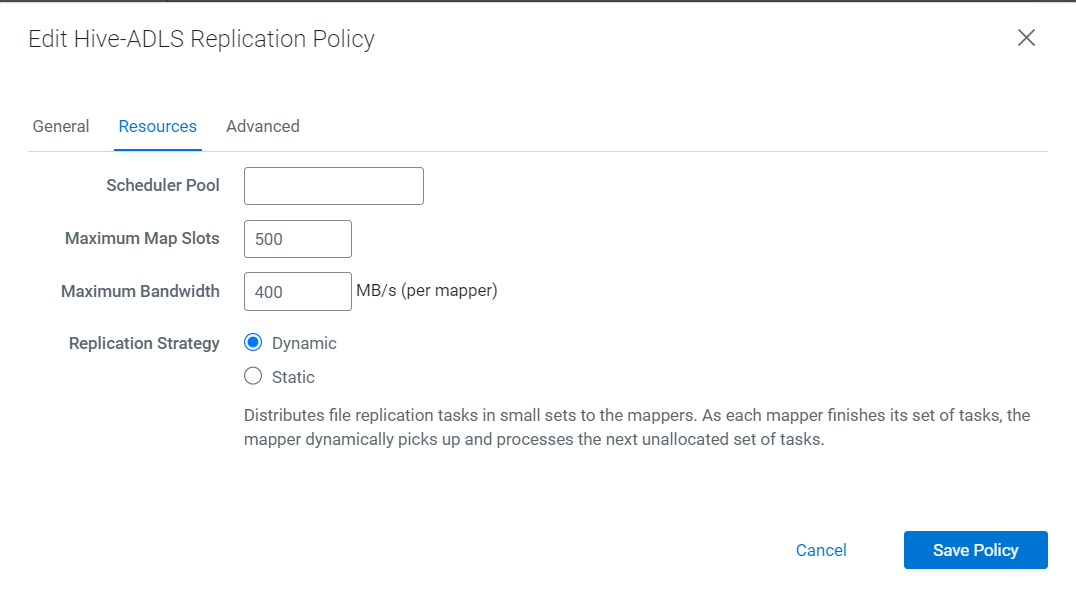
Note:

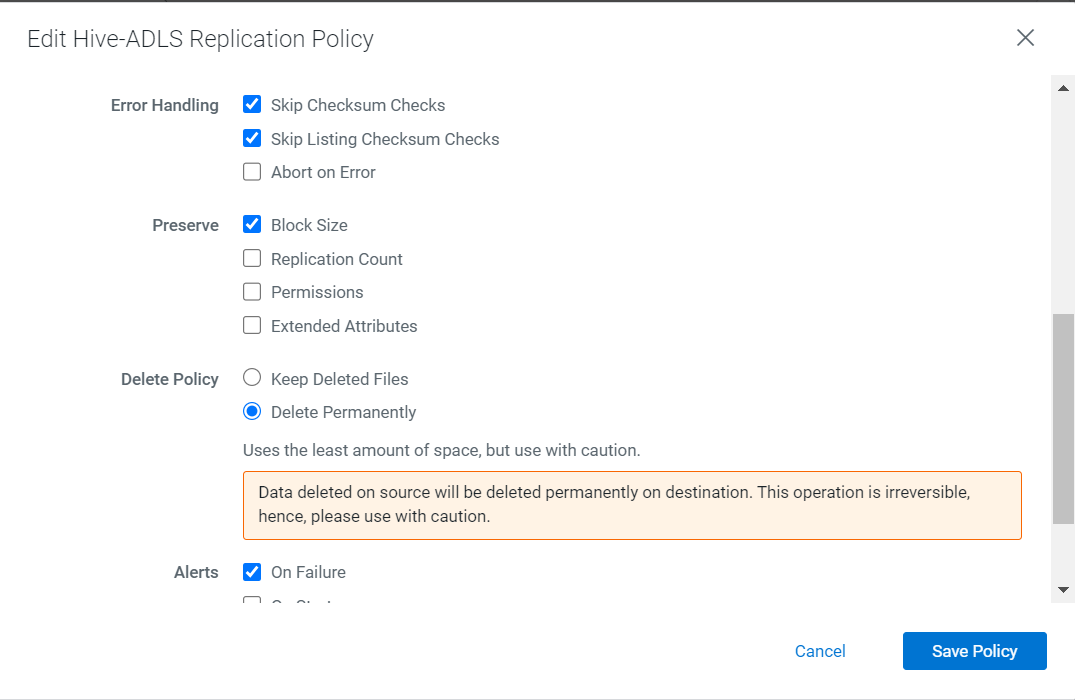
* If you select Hive replication, cloudera path will be appeneded to Landingzone path. With Hive replication, we can load single or multiple tables or whole DB as well
* HDFS replication will be suitable only for one table. You would need to have cloudera source path to give in HDF replication. Please note that the table will be directly copied under the landing zone path you have given. Cloudera path will not be appeneded to landing zone path when you choose HDFS replication.



* Please see the options to be selected for BDR job







# **Sample Code to do Few Tasks:**

## **Databricks Cluster Report:**

root@0603-223137-u9s6n9fp-10-235-176-60:/databricks/driver# cat cluster\_report.py

DATABRICKS\_HOST = "https://adb-6481448659160435.15.azuredatabricks.net"

DATABRICKS\_TOKEN = "d\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"

import pandas as pd

from databricks\_cli.sdk.api\_client import ApiClient

from databricks\_cli.clusters.api import ClusterApi

import datetime

api\_client = ApiClient(

host = DATABRICKS\_HOST,

token= DATABRICKS\_TOKEN

)

clusters\_api = ClusterApi (api\_client)

cluster\_list = clusters\_api.list\_clusters()

cols = ["cluster\_id", "cluster\_name", "init\_scripts"]

list1 = []

list2 = []

list3 = []

for subdict in cluster\_list['clusters']:

list1.append(subdict['cluster\_id'])

list2.append(subdict['cluster\_name'])

try:

list3.append(subdict['init\_scripts'])

except:

list3.append('NA')

print(subdict)

output\_df = pd.DataFrame(list(zip(list1,list2,list3)), columns = cols)

output\_df.to\_csv('/tmp/cluster\_report.csv')

root@0603-223137-u9s6n9fp-10-235-176-60:/databricks/driver#

## **Command to list the jobs which are using policy:**

import requests

import json

# Set the API endpoint and access token

endpoint = "https://adb-1792884946568418.18.azuredatabricks.net/api/2.0"

token = "d\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"

# Set the policy ID

policy\_id = "D96408157A000546"

#policy\_id="D9610A22110002E0"

job\_list=[]

#settings\_list=[]

# Get the jobs associated with the policy ID

response = requests.get(endpoint + "/jobs/list", headers={"Authorization": "Bearer " + token})

jobs = json.loads(response.text)["jobs"]

print(jobs)

# **Information needs to get from new project or team:**

If any new team or project wants Databricks access, please request below details to provide and setup a call with our Hadoop team if any questions.

Is it part of QDR or EIM? If this is a new project team requesting access for Databricks, below steps are to be followed:

* Work with Data governance team associated with this project team, to submit request for new AD groups specific to workspace separately for onsite and offshore teams.
* Data governance team will be responsible to provide brand new AD groups with level of access mentioned to Databricks objects(SQL endpoints, Schemas).

FYI, Governance team contact details are below

EIM - Batchu, Srini <Srinivasarao.Batchu@molinahealthcare.com>; Rao, Ajoy <Ajoy.Rao@molinahealthcare.com>

QDR - Singh, Sandeep [Sandeep.Singh@MolinaHealthCare.Com](mailto:Sandeep.Singh@MolinaHealthCare.Com)

Unfortunately, we do not grant access to individual users. Please work with Data governance team to find the right AD group and then submit a request to become a member of that group.

Data governance details are below.

EIM - Batchu, Srini <Srinivasarao.Batchu@molinahealthcare.com>; Rao, Ajoy <Ajoy.Rao@molinahealthcare.com>

QDR - Singh, Sandeep Sandeep.Singh@MolinaHealthCare.Com

Cancelling this ticket. Please let me know if you have any questions

How do I access Molina Data Lakehouse?  
  
Molina Azure Data Lake can be accessed in a 2-step process.  
Get access to a Citrix VDI (Virtual Machine)  
Get access to Databricks SQL.  
  
To request access to Citrix VDI and Databricks SQL, ask your line manager to send an email to: EIM-DM-DG@molinahealthcare.onmicrosoft.com  
Email should include:  
Resource name, UserID, Molina email ID  
On-shore or off-shore status – This will help the admin decide the right AD group.  
Databricks SQL environment. (UAT or Prod) – This will help the admin decide the right environment.  
Dept name (example: ClaimsOps, MedEcon etc..) – This will help the admin decide the appropriate cluster to give access to.  
\*iServe request process in development, the slides will be updated accordingly.

<https://molinahealthcare.sharepoint.com/sites/EIM-KB/Technical%20Training/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FEIM%2DKB%2FTechnical%20Training%2FShared%20Documents%2FDatabricks%20Training&viewid=b8a64c33%2D0751%2D48e2%2D9c9e%2Def7cf81b7585&OR=Teams%2DHL&CT=1685645228680&clickparams=eyJBcHBOYW1lIjoiVGVhbXMtRGVza3RvcCIsIkFwcFZlcnNpb24iOiIyNy8yMzA1MDEwMDQyMiIsIkhhc0ZlZGVyYXRlZFVzZXIiOmZhbHNlfQ>

We do not grant access to individual users . Please work with Data governance team to find the right AD group and submit a request to become a member of that particular group to get Databricks access.

If you want to get access to new AD groups, please open a new ticket and get approval from Data Governance team.

Data governance team details are below

For EIM - Srini Batchu or Ajoy Rao

For QDr - Sandeep SIngh

[RITM3678007](https://molina.service-now.com/iserve/sc_req_item.do?sys_id=2f0a4e008700b598432685d7cebb3585&sysparm_record_target=sc_req_item&sysparm_catalog=e0d08b13c3330100c8b837659bba8fb4&sysparm_catalog_view=catalog_default)

Since this request is not approved by Data governance team on time and it is aged, we are cancelling this ticket. Please open a new ticket if you still need this access and get approval from Data governance team.

Data governance team details are below

For EIM - Srini Batchu or Ajoy Rao

For QDr - Sandeep SIngh

Reference tickets and Team names and DLs:

Network Operations Center:

Network Operations Center <NOC@MolinaHealthCare.Com>

Citrix:

DL-MOL-INF-COMPUTE-Citrix [Citrix\_Tower@molinahealthcare.com](mailto:Citrix_Tower@molinahealthcare.com)

**New schema/databse creation request:**

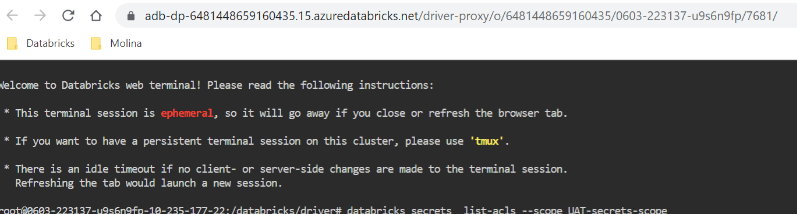
Is it part of QDR or EIM?  
  
Work with Data governance team associated with this project team, to submit request for new schema/databases creation.  
  
Data governance team will be responsible to put a request with all the required details or please update the ticket with proper details such as Database name, database type(l0, l1, etc) , AD groups that require access and access level(ro/rw) and get the approval from Data Governance team. For more details, please contact Data governance team.  
  
  
  
FYI, Governance team contact details are below :  
  
EIM - Batchu, Srini <Srinivasarao.Batchu@molinahealthcare.com>; Rao, Ajoy <Ajoy.Rao@molinahealthcare.com>  
  
QDR - Singh, Sandeep Sandeep.Singh@MolinaHealthCare.Com

**Data refresh:**

We do not have access to copy data from Prod Databricks storage account to UAT Databricks storage account and backend connection itself is not enabled between Prod (scadlsg2datbksprod) and UAT (scadlsg2datbksuat) storage accounts.  
  
Data has to be refreshed in landing zone and then Dataservices team can refresh the table from landing zone.  
  
Please let us know if you want us to refresh the data in landing zone.

# **Configure Databricks CLI:**

* Go to Compute
* Start the “Test\_cluster” if it is not up
* Go to apps and then Click “Web Terminal”
* It launches terminal as below



* Install databricks-cli using below pip command

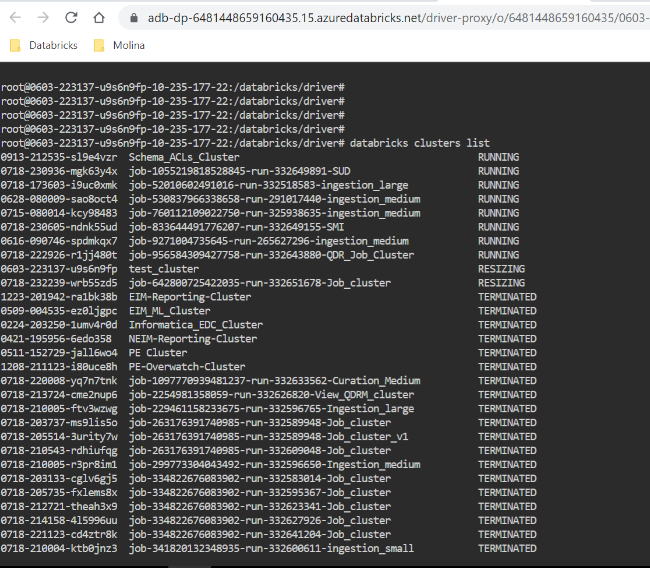
pip install databricks-cli

* Configure token by giving URL and PAT token

Databricks configure --token

* Can run the databricks commands

Example: databricks clusters list



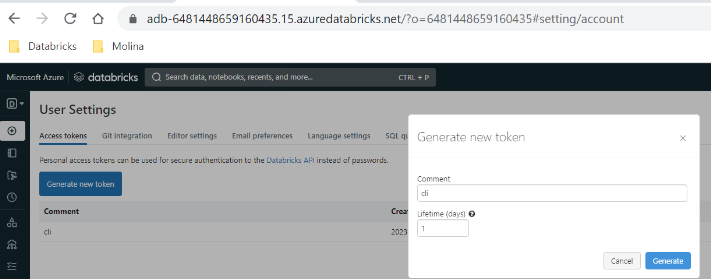
## **Get PAT Token:**

* Login to Databricks Workspace
* Click on profile (Your username)
* Click User Settings
* Click “Generate new Token” under Access Tokens
* Give Comment and select life time

**Note:** Give just 1 or 2 days for token lifetime since profile will go off when the cluster restarts and you would need to reconfigure databricks cli and token .

* Click Generate
* Copy the token and click done.

**Note:** Once you click done, you cannot get the token. You would need to regenerate it.



# **Grant Access to Job cluster policy:**

If any team requests access to Job cluster policy, we would need to grant “Can Use” access to Job cluster policy and “READ” access to the scope configured in policy config.

Reference Ticket - [RITM3690770](https://molina.service-now.com/iserve/sc_req_item.do?sys_id=17805f058704f1d063a2da873cbb35d9&sysparm_view=ess&sysparm_record_target=sc_req_item&sysparm_record_row=2&sysparm_record_rows=2&sysparm_record_list=u_requested_for%3Djavascript%3Ags.getUserID%28%29%5EORopened_by%3Djavascript%3Ags.getUserID%28%29%5Eactive%3Dtrue%5EORDERBYshort_description)

## **Grant access to Policy:**

Go to Compute – Policies

Search with policy name

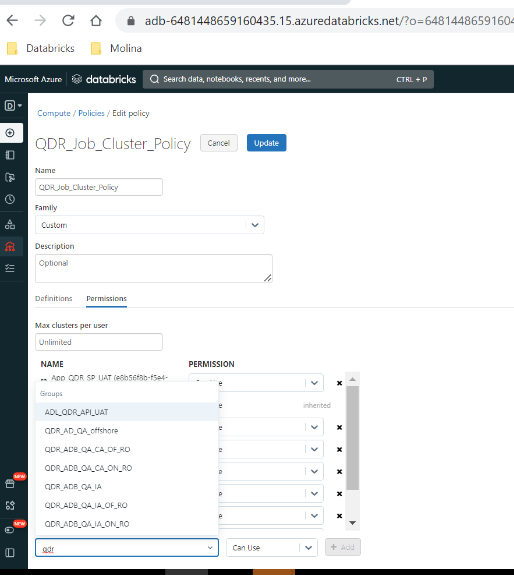
Click Edit

Go to Permissions

Search for AD group

Click Add

Then Click Update



## **Grant access to the scope:**

To grant access to the Scope, you would need to install databricks-cli and configure token in web terminal as shown in the section “Configure Databricks CLI”

Once you installed databricks-cli and configure token, you would need to run databricks secrets commands to grant access.

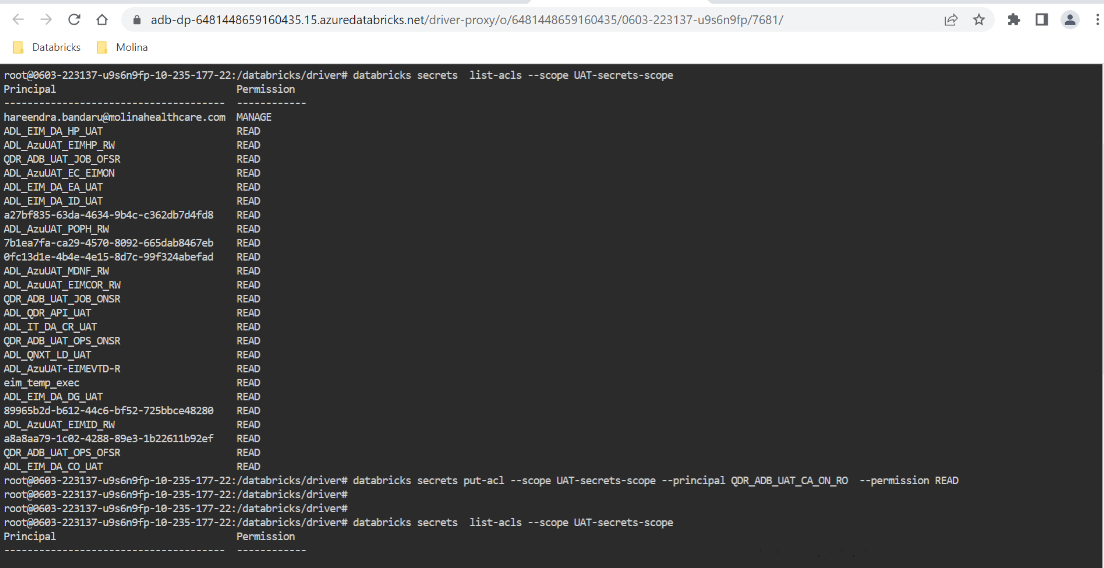
Note: Always grant READ access to the AD groups to the scope. We should not grant access to individual usets.

**Command to grants access to the scope:**

databricks secrets put-acl --scope <ScopeName> --principal <ADGroupName>  --permission READ

**Command to list acls on the scope:**

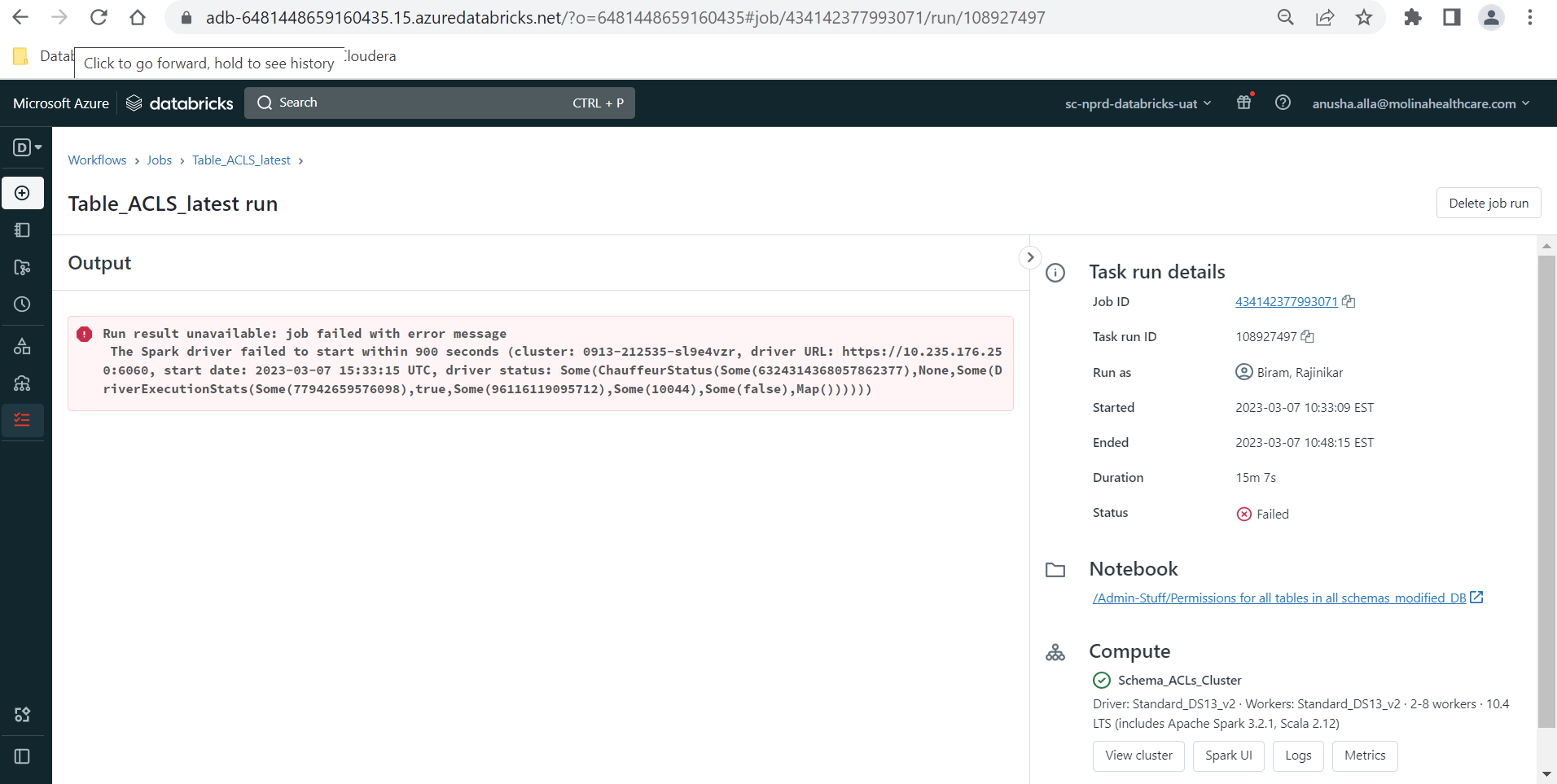
databricks secrets  list-acls --scope <SCopeName>

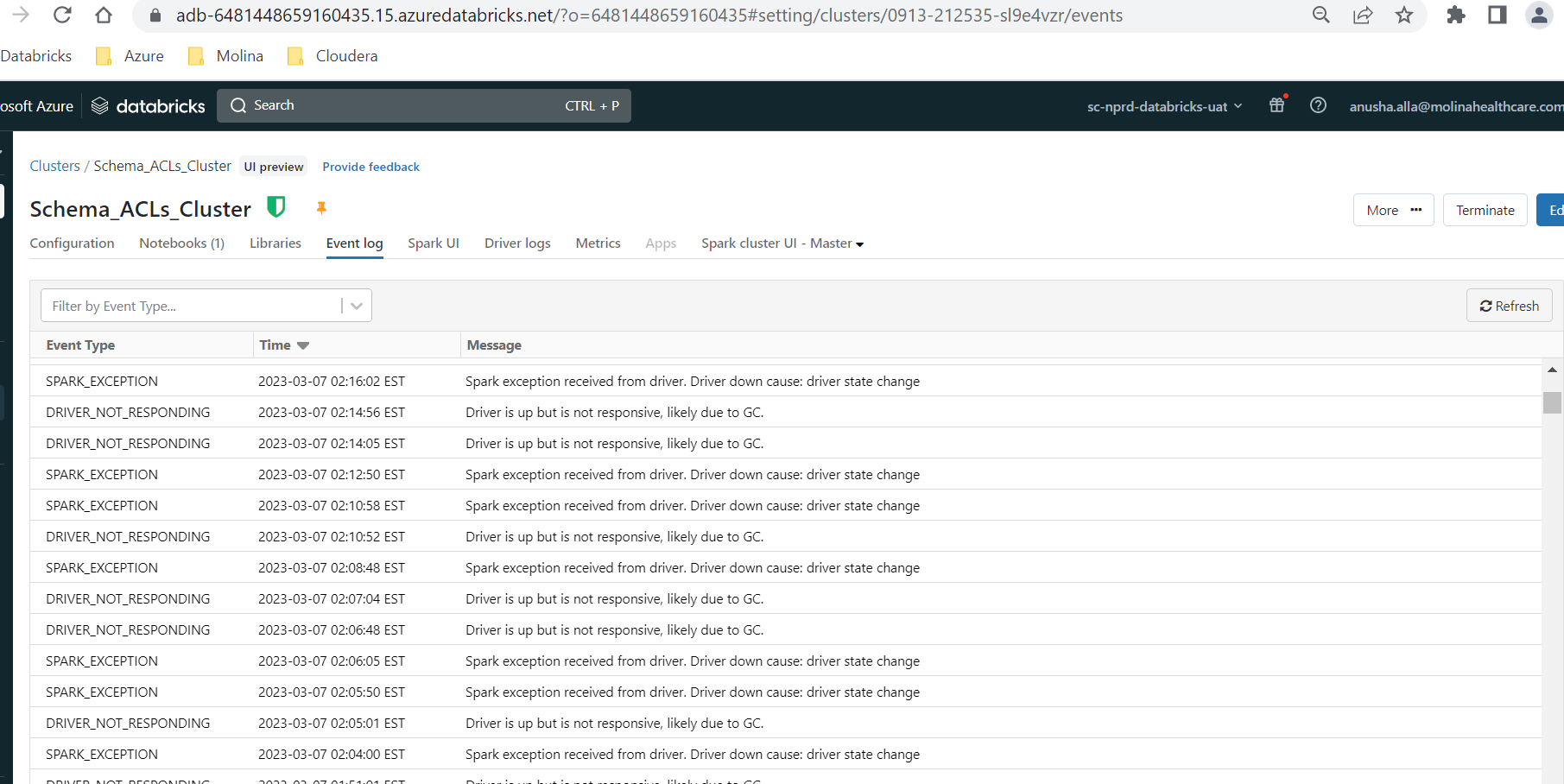


# **Issues and Solutions:**

## **Cluster Issue - Driver is up but is not responsive, likely due to GC:**

Job failed to run on cluster with below error and then when we check event logs of cluster we see this error.



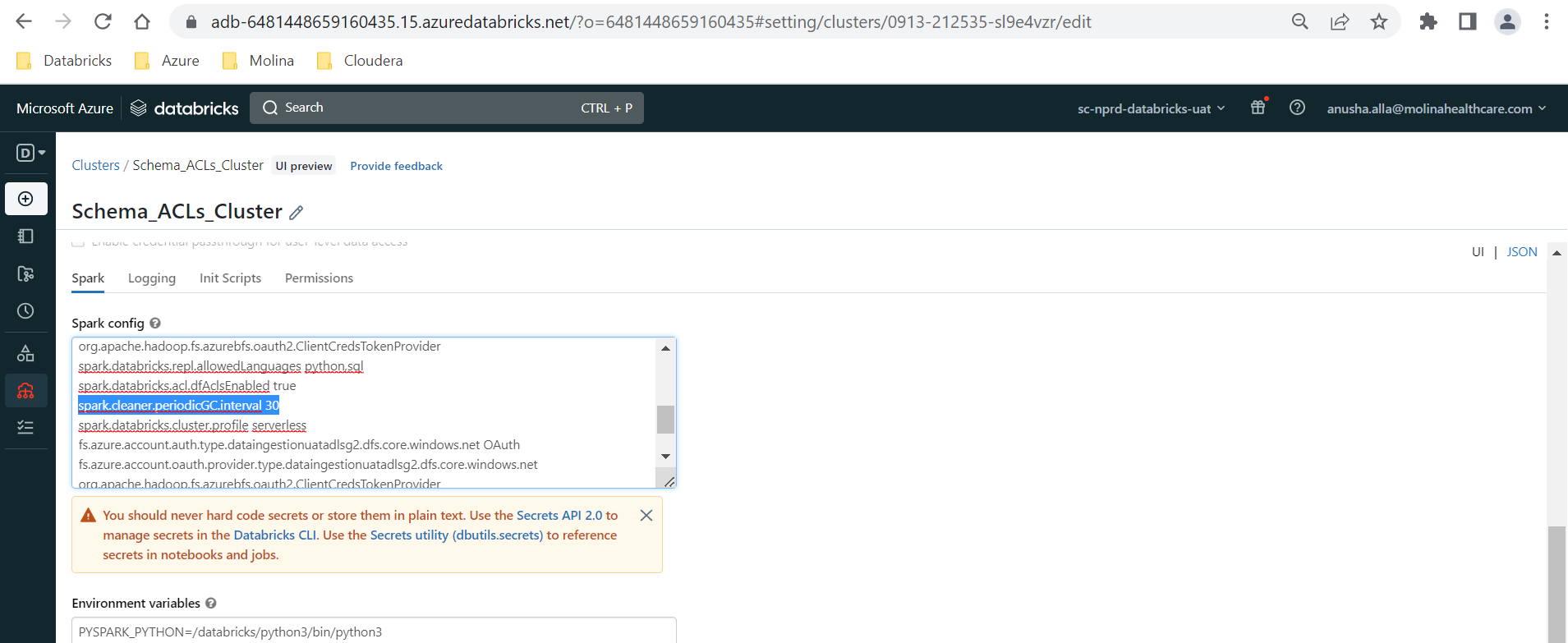


**Solution:**

It is due to driver node out of memory issue. We can try adding Spark config parameter to increase GC cleanup period to 30 mins in the cluster Spark configuration. If it doesn’t work then increasing driver node memory is the solution.

**spark.cleaner.periodicGC.interval 30**

Go to Compute – Click on cluster which you have issue with – Click edit – go to Advanced Options – Add the above parameter in spark Config section.



**Error: org.apache.spark.SparkException: Job aborted.**

Caused by: org.apache.spark.SparkException: Job aborted due to stage failure: Task 235 in stage 4672.0 failed 4 times, most recent failure: Lost task 235.3 in stage 4672.0 (TID 47471) (10.235.17.185 executor 109): ExecutorLostFailure (executor 109 exited caused by one of the running tasks) Reason: Executor heartbeat timed out after 143067 ms

**Issue:**

Job Failed with org.apache.spark.SparkException: Job aborted. At one of the tasks

**Recommendation 1:**

Please try the configurations below at the cluster level.  
   
**spark.databricks.aggressiveWindowDownS 300**  
 **spark.executor.heartbeatInterval 500s**  
 **spark.network.timeout 1000s**

**Recommendation 2**  
   
As we can see Out of Memory error in some of the executors, we request to increase the executor instance memory to the next level or higher.

**Solution**: changed policy from medium to large and rerun. Worked fine

## **Insufficient privileges:**

**Issue:**

Users cannot select on tables due to insufficient privileges on underlying tables/views.

**Error 1:**

[INSUFFICIENT\_PERMISSIONS] Insufficient privileges: User cannot SELECT on table `sourcedata\_db`.`meg\_file` because of permissions on underlying securables: The owner of table `meg\_db`.`file` is different from the owner of table `sourcedata\_db`.`meg\_file`.

If any table/view created out of some table, then make sure source table and target table/view owner are same. If the owner is different, though users have access to the target table, they cannot access it.

In Prod, make sure source and target tables have right service principal as owner.

In Prod, for reporting DB tables, owner should be AD group. Please check with our team or Data governance team before changing the owner.

In UAT, we have a job **“Table\_ACLs\_Owner\_Update”** running twice a day to take care of all the tables ownership. This job changes the owner of all tables to “admins”.

## **Job Failed due to out of memory:**

**Issue:** org.apache.spark.sql.execution.OutOfMemorySparkException: Size of broadcasted table far exceeds estimates and exceeds limit of spark.driver.maxResultSize=4294967296. You can disable broadcasts for this query using set spark.sql.autoBroadcastJoinThreshold=-1

**Solution:** Recommended solutions are

1. Disable Broadcast or,
2. *Use ANALYZE TABLE table\_name COMPUTE STATISTICS.*

Add the below parameter to disable broadcast in the notebook and run the job

spark.sql.autoBroadcastJoinThreshold=-1

<https://kb.databricks.com/sql/bchashjoin-exceeds-bcjointhreshold-oom>

## **Stage Failures:**

### **Worker node lost:**

Task with multiple stages – When it reads from one source it creates one stage, and when it reads from other source then it will create other stage. Also, when it joins/merges, it creates another stage.

If task has 200 stages then it could be doing Shuffle operations – shuffle, join, merge all these are shuffle operations.

If the other stages are completed and taking less than milliseconds and particular failed stage is running for 20 mins because it is loading heavy data. Data is queued.

If task fails, it retries to execute and if it still fails max of 4 times then job gets aborted.

Metrics snapshot takes for every 15 mins.

If max cores used for cluster are 256 and the shuffle partitions are 500 then at max only 256 partitions can be handled, then it will wait for task to complete and create other tasks based on data.

If you want to handle all 500 partitions at a time, then it should have 500 cores available.

For a task, if it handles 128-200MB data then 256 cores/tasks then it can handle max 50 GB data

Scaling up depends on number of pending tasks. For example, in our case it completed 199 out of 200 in 0.2 milli seconds so it didn’t think that it needed to scale up.

Whenever there is shuffle operation, in our case it is join. For example, it joins two tables with 10 records, there could be cross join (one record one table can join with other table of 10 records).

**Suggestions:**

1. Please increase the spark.sql.shuffle.partitions value to 500 in your cluster. You can add below config in the spark tab of the cluster configuration page under the advanced setting  
   spark.sql.shuffle.partitions 500

Incraesing shuffle partitions, it increases number of tasks and then number of stages. So data load segregated among multiple partitions.

1. Instead of using trigger.once() option, you can also go ahead with triggerAvailable option along with maxFilesPerTrigger which will limit the data being read in a single-stream batch and will save the application from overloading. Please refer to below documentation  
   [https://docs.databricks.com/en/structured-streaming/triggers.html#configuring-incremental-batch-processing](https://urldefense.com/v3/__https:/docs.databricks.com/en/structured-streaming/triggers.html*configuring-incremental-batch-processing__;Iw!!DOw_8Fim!JzxWrLWfF42PyP8RUkYNAhk_x2Y_N8jXzAVMB9L06Wlo58mVH_5aFB0h1G15acIEyurJc-keulwwyN_s9DGOEbLQJw$)

Trigger once is processing everything at one time. It will wait for all the data to be available and then shuffle operation will be triggered at one time.

Trigger available operates based on available data. As long as some 10gb available out of 50 GB it performs shuffle operation for 10GB instead of waiting for all the data do heavy operation

### **Relative Path Issue:**

* Relative path issue is occurring on workflows because the naming standards/convections are not meeting the Databricks standards having (:) in timestamps.

**Why it is failing first run and second run is success without making any changes to filesystem/code ?**

There is threshold value for checkpoint to list down all the files under checkpoint folder. Once checkpoint reach to threshold level, checkpoint will run file listing command and read all files under checkpoint path. During the file list reading, if any files are having invalid format file names. It will through relative path error.

Same job run successfully when we do repair run, because during the second run checkpoint bypass files listing under checkpoint and run successfully. If checkpoint run file listing command in second run, job will be failed with same relative path error.

**Analysis on workflows failed with relative path error ?**

If we see all workflows (list of jobs in pervious email) which are failed with relative path error. The files are related to June month. I think ADF team implemented correcting the filenames in later June or July month.

If we can delete or archive the files with invalid filenames. We should be able to fix this issue.

### **Remote RPC client disassociated:**

org.apache.spark.SparkException: Job aborted due to stage failure: Task 28 in stage 102.0 failed 4 times, most recent failure: Lost task 28.3 in stage 102.0 (TID 1195) (10.235.17.197 executor 14): ExecutorLostFailure (executor 14 exited caused by one of the running tasks) Reason: Remote RPC client disassociated. Likely due to containers exceeding thresholds, or network issues. Check driver logs for WARN messages.

Suggestions :

1. Change the groupby to make smaller groups.
2. Job output, such as log output emitted to stdout, is subject to a small size limit. If the total output has a larger size, the run will be canceled and marked as failed."

set the logging level to ERRORval sc = SparkContext.getOrCreate(conf)sc.setLogLevel("ERROR")

1. By default, Databricks/Spark use 200 partitions. For the smaller dataset, that works fine. For the larger dataset, it is too small. The solution is to provide the desired number of partitions as part of the repartition call.

spark\_df = spark\_df.repartition(1000, [KEY\_COLUMNS]

1. Increase resources of the cluster (temporary)
2. add the following cluster configuration options:

spark.network.timeout 10000000

spark.executor.heartbeatInterval 10000000

### **Out Of Memory due to collect() and dbutils.fs.put() functions:**

#### **dbutils.fs.put():**

We have reviewed a few failed runs and we have observed that each job run is failing at dbutils.fs.put command. Just above this command I can see you have a for loop in using which you prepare the JSON data to dump to target path of dbutils.fs.put() operation.  
  
However, this specific error, as far as I know, have specific array limits per OS / JVM. There is also a hard limit for arrays max elements to be under Integer.MAX\_VALUE (2,147,483,647)  
[https://www.thejavaprogrammer.com/requested-array-size-exceeds-vm-limit/](https://urldefense.com/v3/__https:/www.thejavaprogrammer.com/requested-array-size-exceeds-vm-limit/__;!!DOw_8Fim!NYgTVnhwFt6YVp6976jwkkDFDry73YM3N-ENNdO9bGvfPXY0YsFnLBUwwhKoGE89yieyNWU4_adsGyZEcIQaXLyXxMdGyd0SlFVz8w$)

dbutils code runs inside the JVM and the data (string of size 1.7gb+) is in the python memory. So, when this python str is passed to dbutils, it creates an array of characters, and it throws the error.  
  
Hence, I believe that the JSON data the customer put into a destination file using put operation is quite huge resulting in this error.  
  
To avoid this, you can first write a data to a local disk and then copy into a DBFS/Cloud location. For example:  
  
------  
with open('/tmp/destination\_file.txt', 'w') as fp:  
fp.write(str(decrypted\_data))  
------  
and then  
-------  
dbutils.fs.cp('file:/tmp/destination\_file.txt', '/FileStore/destination\_file.txt')  
-------  
PS: Local VM disk files can be accessed by %fs or dbutils by prefixing with "**file:**"t

#### **collect() Function:**

Here is a summary of our meeting, please feel free to correct me if there is anything inaccurate:

1. Failed job URL: [https://adb-3372830949335037.17.azuredatabricks.net/?o=3372830949335037#job/454159322422741/run/1050223886970905](https://urldefense.com/v3/__https:/adb-3372830949335037.17.azuredatabricks.net/?o=3372830949335037*job*454159322422741*run*1050223886970905__;Iy8vLw!!DOw_8Fim!OHk9b5sHhaq9U77P2lPRgCVG0iU4xZ9_wb42vTBlMK90WrPA800cKoCs5FNFpa4lrCgTWeLjzjc-_66fsWd5zZ4aAL4$)
2. This is an OOM issue. We checked the Metrics and Checked Metrics. VMs are hitting memory limits.

Suggestions:

1. Upgrade VM size. 32ds -> 64ds

You are using a lot of 'collect()' syntax, which is taking a lot of memory. This is a very common cause of out of memory issues since it attempts pulling the entire Dataframe into the driver’s memory. The general recommendation is to avoid using it.

If upgrading VM size won’t resolve the issue, please consider code change and avoid using collect()

## **Job Cluster launch Issue**

Job clusters are taking long time around 20-50 mins to start the cluster. We noticed this issue mostly on the clusters where init scripts are present. It is taking lot of time to execute init scripts.

Working with Support to investigate the issue and resolve it.

Thank you for enabling the cluster level logging for the job runs, this gave us lot more insights on the latency in cluster start and we were able to narrow down the issue to init script from the below provided job run urls.

Here is a quick overview of the investigation so far.

* We could see the jobs which are running with min 8 worker nodes, as starting the intiscript on time and 1 out of 8 nodes is having latency in the completion of init script.
* Taking example of one of the job runs: [Run of start of azuredatalake\_l0\_tcs\_load1\_fullload\_daily - Databricks (azuredatabricks.net)](https://urldefense.com/v3/__https:/adb-3372830949335037.17.azuredatabricks.net/?o=3372830949335037*job*867161677868579*run*810548898407554__;Iy8vLw!!DOw_8Fim!MVag5VuQb5SozILllr_tZtih4n6ase3VdR-tzj81CTSJIMZ06EzZ6FixrdanDcaILQzgpQOGl-ryx6bZORtAj92WNcPZ$) which took 19 minutes to complete job took about 16 minutes to start the cluster.
* On checking on the init script logs for the cluster, we could see that we have latency in download of few jars and as init script continues in a linear way, all other jar downloads are still waiting for the task to complete.

Eg:

--2023-09-26 01:57:39--  [https://maven-central.storage-download.googleapis.com/maven2/com/crealytics/spark-excel\_2.12/0.14.0/spark-excel\_2.12-0.14.0.jar](https://urldefense.com/v3/__https:/maven-central.storage-download.googleapis.com/maven2/com/crealytics/spark-excel_2.12/0.14.0/spark-excel_2.12-0.14.0.jar__;!!DOw_8Fim!MVag5VuQb5SozILllr_tZtih4n6ase3VdR-tzj81CTSJIMZ06EzZ6FixrdanDcaILQzgpQOGl-ryx6bZORtAj-Ozi7pe$)

Resolving maven-central.storage-download.googleapis.com (maven-central.storage-download.googleapis.com)... 142.250.114.207, 142.250.113.207, 142.251.116.207, ...

Connecting to maven-central.storage-download.googleapis.com (maven-central.storage-download.googleapis.com)|142.250.114.207|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 6672650 (6.4M) [application/java-archive]

Saving to: ‘spark-excel\_2.12-0.14.0.jar’

     0K .......... .......... .......... .......... ..........  0% 2.34M 3s.  
.

.

2023-09-26 02:12:39 (54.0 MB/s) - Read error at byte 6222242/6672650 (Success). Retrying.

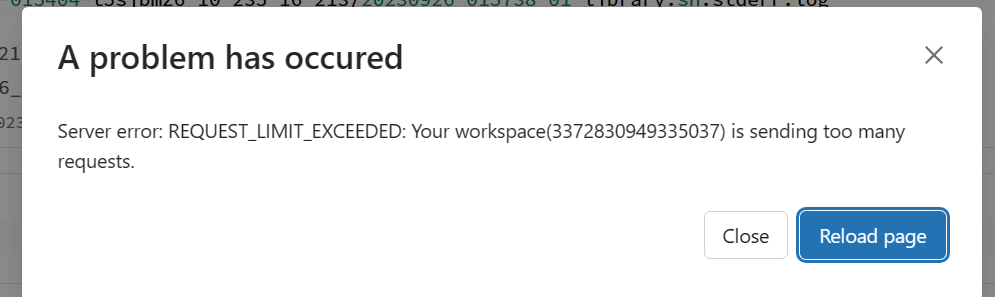
--2023-09-26 02:12:40--  (try: 2)  [https://maven-central.storage-download.googleapis.com/maven2/com/crealytics/spark-excel\_2.12/0.14.0/spark-excel\_2.12-0.14.0.jar](https://urldefense.com/v3/__https:/maven-central.storage-download.googleapis.com/maven2/com/crealytics/spark-excel_2.12/0.14.0/spark-excel_2.12-0.14.0.jar__;!!DOw_8Fim!MVag5VuQb5SozILllr_tZtih4n6ase3VdR-tzj81CTSJIMZ06EzZ6FixrdanDcaILQzgpQOGl-ryx6bZORtAj-Ozi7pe$)

Connecting to maven-central.storage-download.googleapis.com (maven-central.storage-download.googleapis.com)|142.250.114.207|:443... connected.

HTTP request sent, awaiting response... 206 Partial Content

Length: 6672650 (6.4M), 450408 (440K) remaining [application/java-archive]

Saving to: ‘spark-excel\_2.12-0.14.0.jar’

* The pattern of cluster start up latency is seen in all the below jobs, and we
* Also, during peek hours we see that workspace any command execution is throwing below error and notifications which could be a contributing factor to this cluster start up latency.  
  
* Also we have the customer workspace has Unity Catalog enabled and current job cluster we are using is Non-UC enabled cluster ( running on DBR 10.4).

Next Steps plans:

1. Meet SLA of current jobs: as the cluster are not UC enable cluster, we are thinking of approaching mitigation my referring to downloaded jars in dbfs instead of Maven Repo.  
   -- this will avoid the number of outgoing requests from workspace to Maven , and also saves network bandwidth.  
   --this will improve performance in cluster start up as 50% of the task of downloading the jar and then installing on the cluster will be cut down.  
   **Action Pending**: [Rajinikar](mailto:Rajinikar.Biram@MolinaHealthCare.Com), to identify the cluster jars needed, download them to DBFS, and change the init script pointing to maven to point to dbfs jar locations.
2. UC feasible jar installation: As this workspace is UC enabled and any UC enable cluster will not be able to access dbfs, need check if there are ways to install jars as cluster libraries. **Challenges:** Jar files installation from Volume we can try, if volume is not external location volume or dbfs dependent.   
   **Action Pending**: [@Poornima](mailto:pputtaswamy@microsoft.com) to check on the feasibility of jars on UC cluster.
3. MSFT to check on the above-mentioned throttling error message- throttling during Sep 27th AM 8:30 PST – 9:30 PST.

## **Max out the Task Runs in workspace**

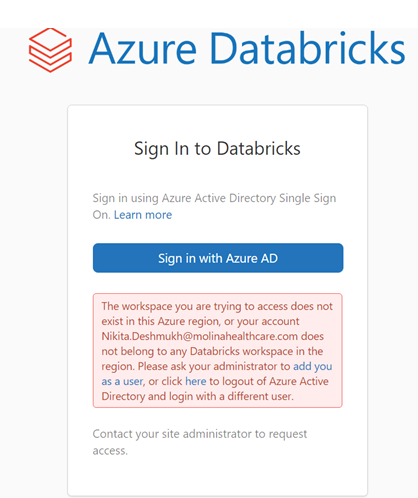
The active max runs within the workspace are 1000. We can increase max upto 2000 but with the help of Support team.

<https://kb.databricks.com/jobs/job-rate-limit>

[Create and run Databricks Jobs | Databricks on AWS](https://docs.databricks.com/en/workflows/jobs/create-run-jobs.html)

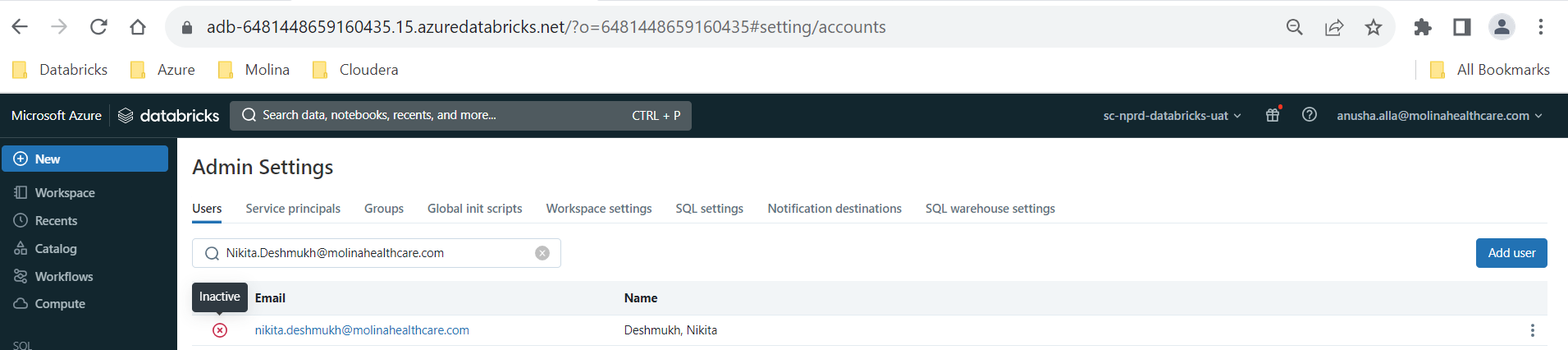
## **Inactive User Issue:**

Molina ID got deactivated and due to that Databricks access got revoked. Once the user is active still facing issue with accessing Databricks.



**Solution:**

If you check the user availability in workspace, you would see user is inactive in the workspace as shown in below screenshot.



## **Update node type in Policies not affecting in some existing workflows**

After updating compute policies to use latest node type, it is not taking affect in existing workflows automatically.

For the single job compute workflows, we have to manually change the configuration or need reselect the policy. But, for the workflows have more than one job compute are not updating with latest node type though we try manually. Getting below error.

## **ColumnMappingUnsupportedException:**

Reading tables enabled with column mapping works only on 10.2 DB Runtime and above. RQES engineering cluster is currently running on 9.1 DB Runtime. We would like to test it on 10.4 DBR and see if this error resolves. Please let me know how you want to proceed.

KB article for reference: <https://learn.microsoft.com/en-us/azure/databricks/delta/delta-column-mapping>

## **Parquet does not support timestamp.See HIVE-6384:**

**Solution:** Created a cluster with below spark parameters and storage credentials then dropped the table using that cluster

spark.sql.hive.metastore.version 1.2.1

spark.sql.hive.metastore.jars maven

**MS recommendation:**

Engineering team recommended the use of below doc to perform the action  
[https://kb.databricks.com/en\_US/metastore/parquet-timestamp-requires-msver12](https://urldefense.com/v3/__https:/kb.databricks.com/en_US/metastore/parquet-timestamp-requires-msver12__;!!DOw_8Fim!LlK7XcM2-KK8CzpGsluj_AeHMyu0sy3bWyM0_vRwdD0cf4ytSnPEUDNMZhBxJb8_FEFqob51Xb8H-qDeYR2gHvz-BAX9yLQz$)  
  
You can do this by adding the following settings to the cluster’s Spark config  
create single node cluster with Databricks Runtime 6.6 and below

* spark.sql.hive.metastore.version 1.2.1
* spark.sql.hive.metastore.jars builtin

**or** single node cluster with Databricks Runtime 7.0 and above

* spark.sql.hive.metastore.jars /dbfs <path-to-downloaded-jars>
* Spark.sql.hive.metastore.version 1.2.1

**Notes:**

1.Just want to make sure that it is not going to upgrade the hive metastore in the workspace or it is not going to affect any databricks objects right?  
This is at cluster level, so it won't impact the workspace  
  
2.Is there any way to check the hive metastore client version on the cluster?

In a python notebook you could run   
  
metastore\_version = spark.sql("SELECT VERSION() AS hive\_metastore\_version")  
metastore\_version.show()

3.Once we update the version of client hive metastore on particular cluster, I should be able to access the table or drop it right?

Yes

4.Is there any way to update hive metastore client on all the all-purpose or job clusters in the workspace if we want? Is it recommended from Databricks ? how it impacts?

Hive 1.2.0 and 1.2.1 are not the built-in metastore on Databricks Runtime 7.0 and above. It can be updated if you need or if you want to use Hive 1.2.0 or 1.2.1 in this case is to use timestamp in parquet

## **Bad Request Error while uploading file and creating table using SQLwarehouse:**

A screenshot of a computer

Description automatically generated  
  
Cause:  
Platform issue with SQL warehouse  
  
Resolution:  
Bug fix ETA Feb 5th 2024.  
ES-954109

# **Service Principals, PAT Tokens and Keyvault Secrets:**

Keyvaults Information:

Scopes Information:

Dev\_secrets\_scope

UAT-secrets-scope

# **Third-Party Teams details:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Team | iServe Group Name | DL | Reference RITM or INC | iServe Request Catalog Form |
| Azure Ops | PE-Solution Engineering-Azure Operations Support |  | INC3197244 |  |
| Ingestion Team | EIM-Data Ingestions-LZ |  |  | [Data Ingestion (Attunity,DataFactory) - Request | Molina Service Management Platform (service-now.com)](https://molina.service-now.com/now/nav/ui/classic/params/target/com.glideapp.servicecatalog_cat_item_view.do) |

# **DBUTILS Commands:**

## **Cleanup the files/subfolders under given folder:**

PATH = "abfss://outbound-zone@scadlsg2datbksobzoneprod.dfs.core.windows.net/eim\_dm/mhil\_uic\_dscc/"

for i in dbutils.fs.ls(PATH):

    dbutils.fs.rm(i[0],True)

# **Storage Accounts and SPNs:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Environment** | **Storage Account** | **SPN** | **Role** | **Keyvault Secrets** |
| DEV | dataingestiondevadlsg2 | SC-NPRD-Datbks-DEV-Lndgzn-SPN | Storage Blob Data Contributor | SPN-Lndgzn-Client-ID  SPN-Lndgzn-secret |
| UAT | dataingestionuatadlsg2 | SC-NPRD-Datbks-UAT-RO-SPN | Storage Blob Data Reader | RO-SPN-Client-ID  RO-SPN-Secret |
| UAT | dataingestionuatadlsg2 | SC-NPRD-Datbks-UAT-Lndgzn-SPN | Storage Blob Data Contributor | SPN-Lndgzn-Client-ID  SPN-Lndgzn-secret |
| PROD | dataingestionprdadlsg2 | **SC-PROD-Databricks-RO-SPN** | Storage Blob Data Reader | RO-SPN-Client-ID  RO-SPN-Secret |
| PROD | dataingestionprdadlsg2 | **SC-PROD-Databricks-Lndgzn-SPN** | Storage Blob Data Contributor | SPN-Lndgzn-Client-ID  SPN-Lndgzn-secret |
|  |  |  |  |  |
| DEV | scadlsg2datbksdev |  |  |  |
| UAT | scadlsg2datbksuat | **SC-NPRD-Datbks-UAT-ADLS-SPN** |  | SPN-ADLS-Client-ID  SPN-ADLS-secret |
| PROD | scadlsg2datbksprod |  |  |  |
|  |  |  |  |  |
| DEV |  |  |  |  |
| UAT | scadlsg2datbksobzoneuat | **SC-NPRD-Datbks-UAT-ADLS-SPN** |  | SPN-ADLS-Client-ID  SPN-ADLS-secret |
| UAT | scadlsg2datbksobzoneuat | **SC-NPRD-Datbks-UAT-ADLS-SPN** |  | SPN-ADLS-Client-ID  SPN-ADLS-secret |

# **Script to delete inactive users:**

* Run the below code in a notebook to get inactive users from the workspace/account

**Note:** Give the right workspace URL and PAT token

%sh

sudo apt-get install jq

curl --request GET \

'https://adb-1792884946568418.18.azuredatabricks.net/api/2.0/preview/scim/v2/Users?filter==active%20eq%20False' \

--header 'Authorization: Bearer d\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*' \

--header 'Content-Type: application/json' | jq . >> /tmp/inactiveusers.txt

cat /tmp/inactiveusers.txt | grep -i "userName"| awk -F":" '{print $2,","}' |sed '$ s/.$//'

* Run the below code in a notebook to remove the inactive users.

**Note:** Please pass the list of inactive users which you got from the above code to the variable “inactive\_users\_list” in the below code

import requests

ws\_url = "https://adb-1792884946568418.18.azuredatabricks.net"

list\_users\_url = ws\_url + "/api/2.0/preview/scim/v2/Users"

inactive\_users\_list = ["abhishek.agarwal@molinahealthcare.com"]

#inactive\_users\_list = [i.lower() for in in inactive\_users\_list]

headers = {

'Authorization': 'Bearer d\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*'

}

ws\_users\_list = requests.request("GET", list\_users\_url, headers=headers).json()

delete\_user\_ids = []

for user in ws\_users\_list['Resources']:

if user['userName'] in inactive\_users\_list:

delete\_user\_ids.append(user['id'])

for delete\_user\_id in delete\_user\_ids:

try:

response= response = requests.request("DELETE", list\_users\_url + "/" + str(delete\_user\_id), headers=headers)

print(response.status\_code)

except Exception as ex:

print(ex)

# **Database access Request instructions:**

To request access to any Database/security group, following steps needed in sequence

* Tower leads (Reporting Director or above ) compiling user list and sending it for access request to DG through template.
  + This email will be attached to Access request to document manager approval.
  + **EIM - DG - Data Governance\_Team** [**EIM-DM-DG@MolinaHealthcare.onmicrosoft.com**](mailto:EIM-DM-DG@MolinaHealthcare.onmicrosoft.com)
* DG team review the list and then submit access request (attaching above email and user list and AD groups)
* Request goes to DG lead for review and approval
* After DG lead approval, goes to AD group owner PE lead for approval
* After Security group owner approval, goes to idm for fulfillment

# **Moving a schema from one ADLS location to Other:**

1. **Set up your ADLS configurations:** You'll need to set up configurations to connect to your ADLS.  You generally do this using the dbutils utility in a Databricks notebook:

dbutils.fs.configure(

  azure\_account\_key="<your-adls-account-name>.[dfs.core.windows.net](https://urldefense.com/v3/__http:/dfs.core.windows.net__;!!DOw_8Fim!KZeNj9tq4w2nV9jQZV2IIphkw2V9xMeXA8KcxCh4lDBQ9OiZbUiCCJBPt6nAqkGrqnAZmQO0IglteQxVrFJw6jebO_XJxQ$)=<access-key>"

)

2. **List the tables in the schema:** You may need to list all the tables that belong to that schema (database)

tables = spark.sql(f"SHOW TABLES IN {database\_name}").select("tableName").collect()

3. **Copy the tables to the new ADLS location:**  Loop through the tables and copy them to the new ADLS path.

new\_adls\_path = "abfss://[new-container@new-account.dfs.core.windows.net/new-path/](https://urldefense.com/v3/__http:/new-container@new-account.dfs.core.windows.net/new-path/__;!!DOw_8Fim!KZeNj9tq4w2nV9jQZV2IIphkw2V9xMeXA8KcxCh4lDBQ9OiZbUiCCJBPt6nAqkGrqnAZmQO0IglteQxVrFJw6jecLv9_pg$)"

for table in tables:

    table\_name = table.tableName

    df = spark.table(f"{database\_name}.{table\_name}")

    df.write.mode("overwrite").parquet(f"{new\_adls\_path}/{table\_name}")

4. **Update table locations:** Once you have copied the data, you'll likely want to update the metadata in your Spark/Hive catalog to point to the new location.

for table in tables:

    table\_name = table.tableName

    spark.sql(f"ALTER TABLE {database\_name}.{table\_name} SET LOCATION '{new\_adls\_path}/{table\_name}'")

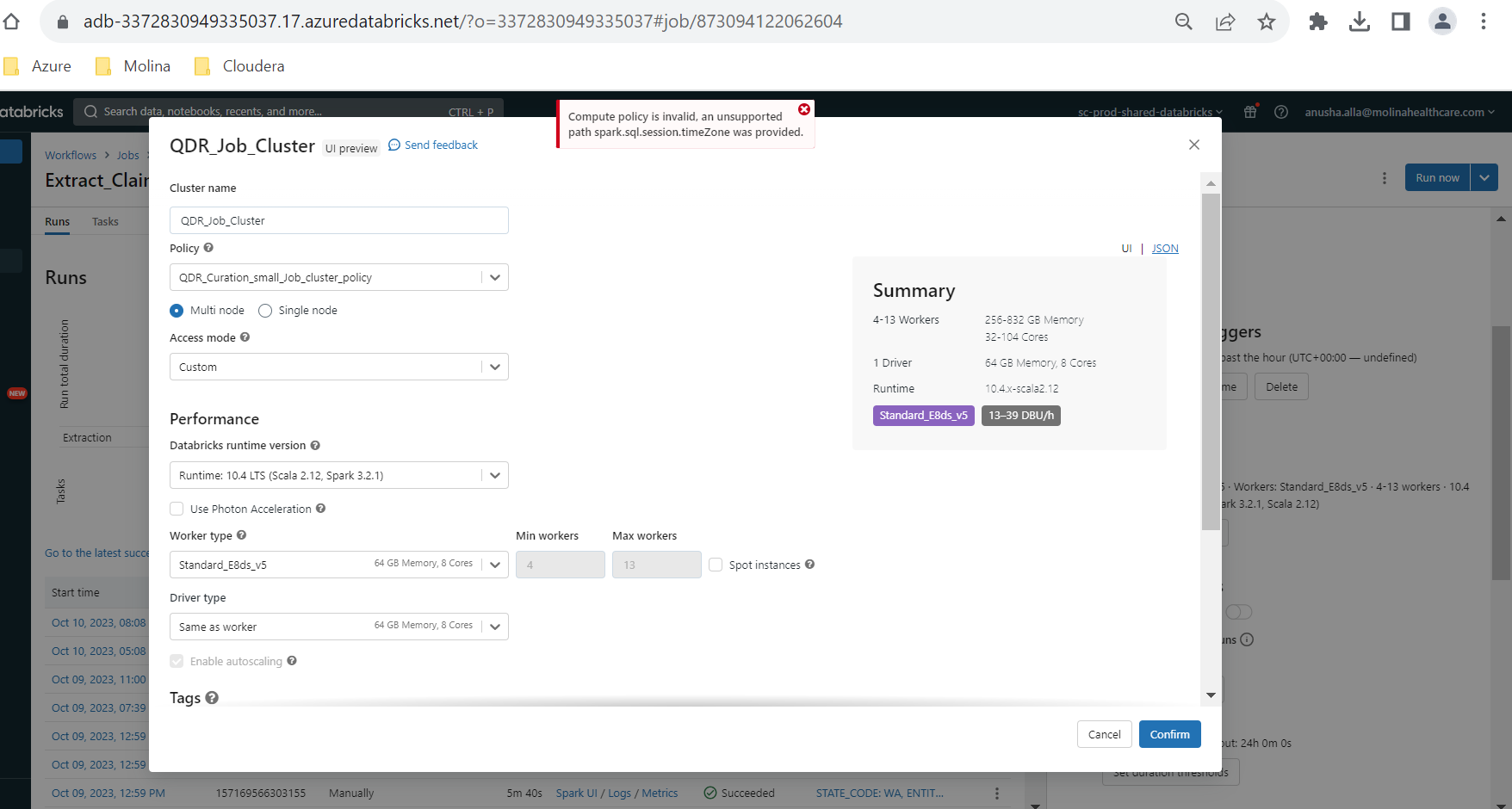
5.  **Delete the old data (Optional):** If you want to delete the old data, you can do so.  Make sure you've verified that the new data is in place and accessible before you do this.

for table in tables:

    table\_name = table.tableName

    dbutils.fs.rm(f"{old\_adls\_path}/{table\_name}", recurse=True)

This is a simplified example, and your actual implementation will vary based on specific requirements and configurations.  Also, before running any operations that modify or delete data, make sure you have adequate backs and have thoroughly tested the process on a subset of data.



# **Add Service Principal to workspace:**

# **Generate PAT Token for Service principal:**

Login to workspace and add the service principal using above section. Run the below code in a notebook. Make sure you add Service Principal application id and application secret and enter correct subscription id.

|  |
| --- |
| import requests  # from wsSecret import KeyVault  #keyvault = KeyVault('https://sc-databricks-kv-uat.vault.azure.net/')  # This is the tenant's login application id, in this case the ID  # for Azure Databricks  login\_application\_id = '2ff814a6-3304-4ab8-85cb-cd0e6f879c1d'  # SPN Info:  # NOTE: Set the values for your service principal  application\_id = "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"  application\_secret = "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"  subscription\_id = "757a3960-ddbd-41d0-a955-3fb9170471ea"  tenant\_id = "5e625f8d-0b53-4f56-9e46-19fa14bb2e5c"  print('subscription\_id:', subscription\_id)  print('tenant\_id:', tenant\_id)  print('application\_id:', application\_id)  # print('application\_secret:', application\_secret)  response = requests.get(      f'https://login.microsoftonline.com/{tenant\_id}/oauth2/token',      headers={'Content-Type': 'application/x-www-form-urlencoded'},      data=f'grant\_type=client\_credentials&client\_id={application\_id}&resource={login\_application\_id}&client\_secret={application\_secret}'  )  print ('response.status\_code:', response.status\_code)  if response.status\_code == 200:      resp = response.json()      access\_token = resp['access\_token']      print('====>> access\_token:', access\_token)  management\_resource\_endpoint = 'https://management.core.windows.net/'  response = requests.get(      f'https://login.microsoftonline.com/{tenant\_id}/oauth2/token',      headers={'Content-Type': 'application/x-www-form-urlencoded'},      data=f'grant\_type=client\_credentials&client\_id={application\_id}&resource={management\_resource\_endpoint}&client\_secret={application\_secret}'  )  if response.status\_code == 200:      resp = response.json()      management\_access\_token = resp['access\_token']      print('====>> management\_access\_token:', management\_access\_token)  DATABRICKS\_HOST = "adb-6481448659160435.15.azuredatabricks.net"  response = requests.post(      'https://%s/api/2.0/token/create' % (DATABRICKS\_HOST),      headers={'Authorization': 'Bearer %s' % access\_token,               'X-Databricks-Azure-SP-Management-Token': management\_access\_token               },      json={          "comment": "This is a test token",          "lifetime\_seconds": 150        }    )  print('gen token response:', response)  print('====>> token info:', response.text) |

# **Init Scripts to refer files in Volumes in Unity Catalog volumes:**

* we can use jars in UC volumes, it shouldn't be a problem.
* copy your libraries to /Volumes/<catalog>/<schema>/<volume>/<path\_to\_library\_file> then refer the paths in init script

Please refer below public docs:

[https://learn.microsoft.com/en-us/azure/databricks/init-scripts/referencing-files#can-i-reference-files-in-unity-catalog-volumes-from-init-scripts](https://urldefense.com/v3/__https:/learn.microsoft.com/en-us/azure/databricks/init-scripts/referencing-files*can-i-reference-files-in-unity-catalog-volumes-from-init-scripts__;Iw!!DOw_8Fim!MV7A3gatUPlLmINBRg8xY9DD3h0XOo4ZYKTzaDG98bMWETEEXO95w4GAI9KYAMTiBUR3-wOClcoh-UQkz5OTpYiO-kr3bQ$)

[https://learn.microsoft.com/en-us/azure/databricks/compute/compatibility#best-practices](https://urldefense.com/v3/__https:/learn.microsoft.com/en-us/azure/databricks/compute/compatibility*best-practices__;Iw!!DOw_8Fim!MV7A3gatUPlLmINBRg8xY9DD3h0XOo4ZYKTzaDG98bMWETEEXO95w4GAI9KYAMTiBUR3-wOClcoh-UQkz5OTpYg2x9-lmg$)

Pip packages install from Volumes [https://learn.microsoft.com/en-us/azure/databricks/libraries/notebooks-python-libraries#--install-a-package-from-a-volume-with-pip](https://urldefense.com/v3/__https:/learn.microsoft.com/en-us/azure/databricks/libraries/notebooks-python-libraries*--install-a-package-from-a-volume-with-pip__;Iw!!DOw_8Fim!MV7A3gatUPlLmINBRg8xY9DD3h0XOo4ZYKTzaDG98bMWETEEXO95w4GAI9KYAMTiBUR3-wOClcoh-UQkz5OTpYjTRan99A$)

Init scripts should be stored in Unity Catalog volumes if using compute with shared or assigned access mode. Workspace files should be used for init scripts if using compute with no-isolation shared access mode.

No-isolation shared access mode does not support volumes, but uses the same identity assignment as shared access mode.

# **Accessing storage account from Cluster/Job policy:**

SPN - APP registration AAD team

It has client id and secret (username and password) - Keyvault as secrets

RO - SC-NPRD-Datbks-UAT-RO-SPN

Scadlsg2datbksuat: (storage account)

Azure Ops - Read only access to SC-NPRD-Datbks-UAT-RO-SPN

Rw - SC-NPRD-Datbks-UAT-ADLS-SPN

outbo

repo

spn -ro and rw

Read the secrets from KV - Scope

keyvault backed rw scope UAT - linked UAT keyvault - fetch secrets

SCope 1 - UAT keyvault

Use this scope to fetch any secrets from the keyvault - eim, HIe, Ramp (AD groups)

scope - Read access to the groups and SPNs

Scope creation:

Commands:

pip install databricks-cli

databricks configure --token

-- Give workspace url and PAT token

databricks secrets list-scopes

databricks secrets list-acls --scope <scopename>

databricks secrets put-acl --scope <scopename> --principal <SPN/AD group> --permission READ

# **Inbound and Outbound Applications:**

|  |  |  |
| --- | --- | --- |
| **Inbound Application** | **User Case** | **Service Principal** |
| **Informatica** |  |  |
| **Power BI** |  |  |
| **Toad (Inactive)** |  |  |
| **SSRS** |  |  |
| **Autosys** |  |  |
| **Diplomat** |  |  |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **Outbound Application** | **User Case** | **Service Principal** |
| **AAS Cube** |  |  |
| **Reporting Applications** |  |  |
| **Diplomat** |  |  |
|  |  |  |
|  |  |  |

**PROD Service Principals:**

|  |  |  |
| --- | --- | --- |
| **Service Principal** | **Application** | **Use Case** |
| APP\_HIE\_PROD | HIE | Deploying and Running the Workflows |
| App\_QDR\_SP\_PROD | QDRM | Deploying and Running Workflows  Scheduling workflows from Autosys |
| EIM\_Mosaic\_Databricks\_PROD | Mosaic | Used in Java Applications to connect to DBSQL (external) |
| SC-PROD-Databricks-Job-SPN | Not in Use |  |
| SC-PROD-Datbks-INFM-SPN | Informatica | Running queries on interactive cluster |
| SP\_ADLAzu\_EIM\_DevOps\_PROD | EIM | Deploying and Running workflows  Scheduling workflows from Autosys |
| SPN-ADL-QDR-API-PROD | QDR | For downstream applications accessing QDR data through API calls |
| SPN-ADL-QDR-MRF-PROD | MRF | Deploying and Running workflows |
| SPN-ADL-QDR-RAMP-PROD | RAMP | Deploying and Running workflows |
| SPN-ADL-QNXTLOAD-PROD | QNXT | Running queries on SQL endpoints connecting through jdbc |
| SPN-ADL-TABLEAU-PROD | Tableau | Running queries on SQL endpoints connecting through jdbc |
| SPN\_ADL\_DG\_Metric\_PRD | DG | Running sql queries and workflows for Data Governance activities using SQL end point and Job clusters |
| SPN\_ADL\_EIM\_DA\_CO\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |
| SPN\_ADL\_EIM\_DA\_DG\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |
| SPN\_ADL\_EIM\_DA\_EA\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |
| SPN\_ADL\_EIM\_DA\_HP\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |
| SPN\_ADL\_EIM\_DA\_ID\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |
| SPN\_ADL\_ENC\_NE\_CO\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |
| SPN\_ADL\_IT\_DA\_CR\_PRD | EIM Reporting | Deploying and Running workflows & queries related to specific reports |