Azure Databricks Assignment 6

Pratik Wani

- Notes

PAGE NO.:

- to store data without performing analysis on data set hirarchical namespace disabled.
- To set it to Datalake Gen2 enabled hirarchical namespece option.

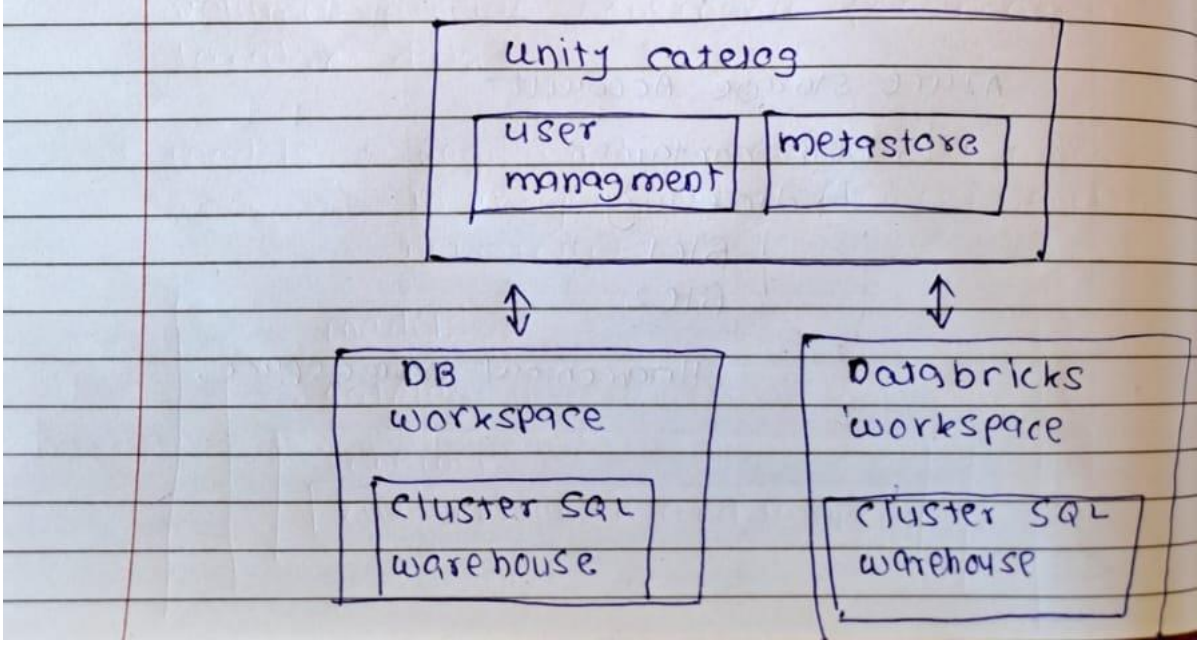
* stage of processing big data

Azure databricks

Day 6

* unity catelog :-

→ Provides centralical access control, auditing, linkage and data discovery capabilities



PAGE NO.:

* Key features :-

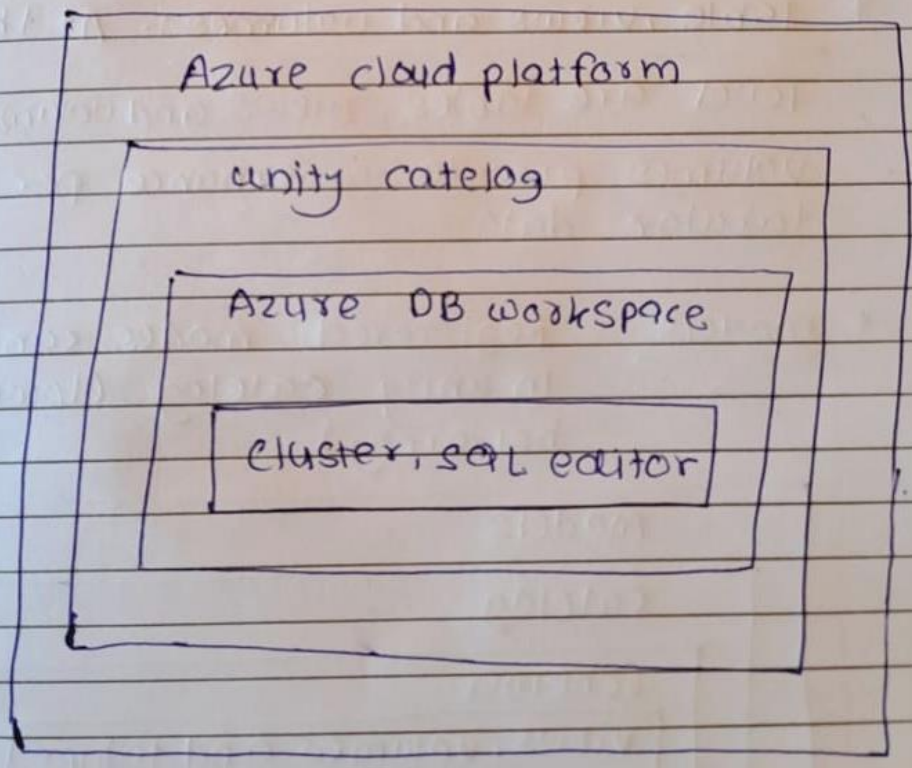
1) Standards :- Complaint security model based on ANSI SQL allows administration to grant permissions in their existing data lake

2) Define once, secure everywhere

3) Built in Auditing and lineage - user level logs

4) data discovery

5) System tables



PAGE NO.:

* Unity Catelog object model:-

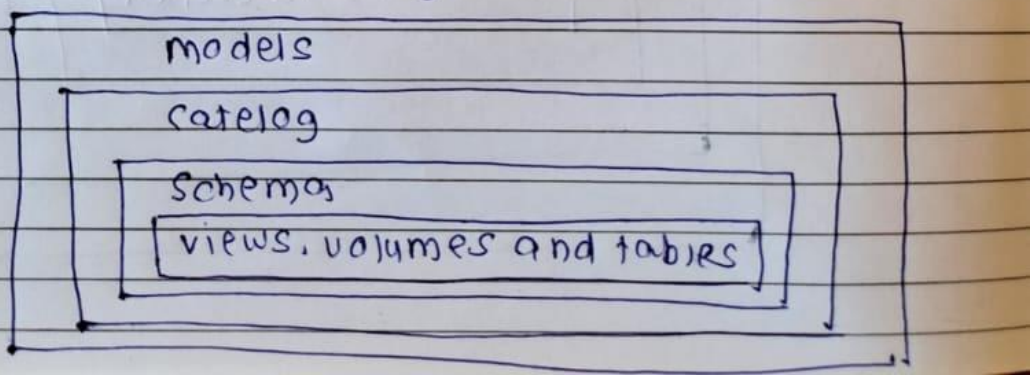
* metastore :: top model container for metastore
Each metastore exposes a three-level namespace (catelog.schema.table)

* catalog :- first layer of object hierarchy used to organize data assess

* schema :- Databases / Schemas are second layer of object hierarchy; contains table and virus

* Table, views and volumes :- At the lowest level are tables, views and volumes
volumes provide governance for non-tabular data

* models : Registered modes can be managed in unity catelog (lowest level in heirarchy)



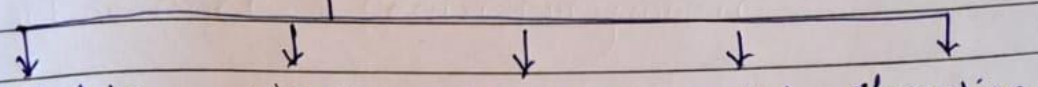
Metastore

|

catalog

|

Schema



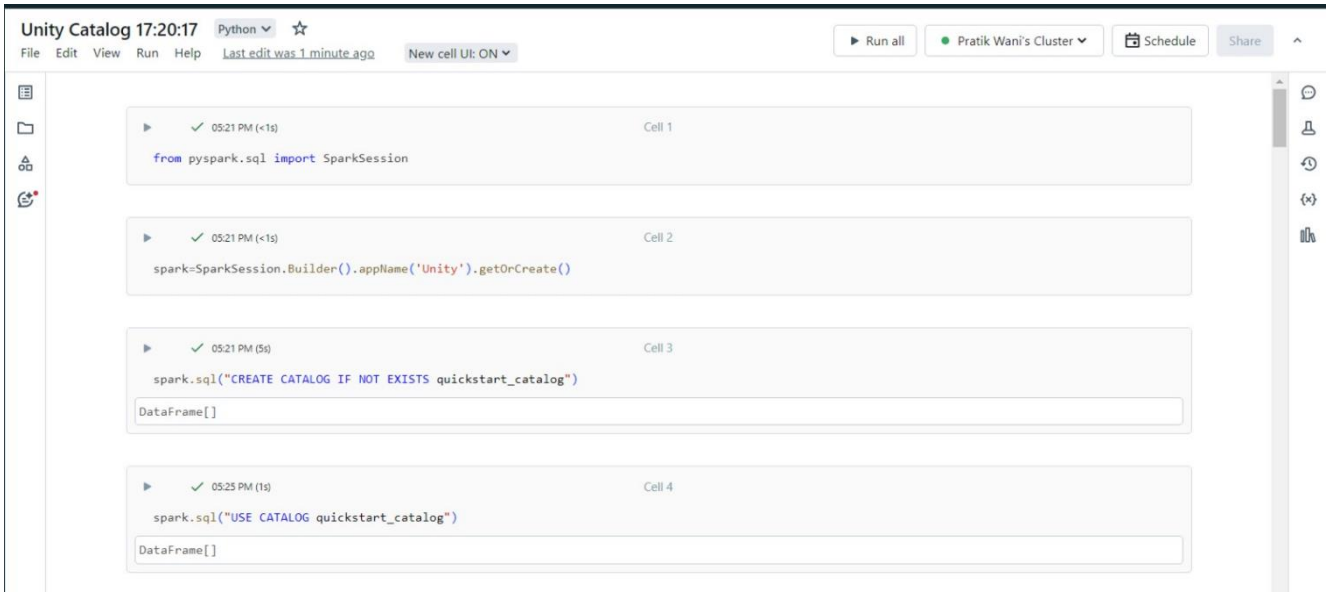
* Azure databricks account admins
→ create one metastore for each region
→ in which they operate and assign it to DB workspace

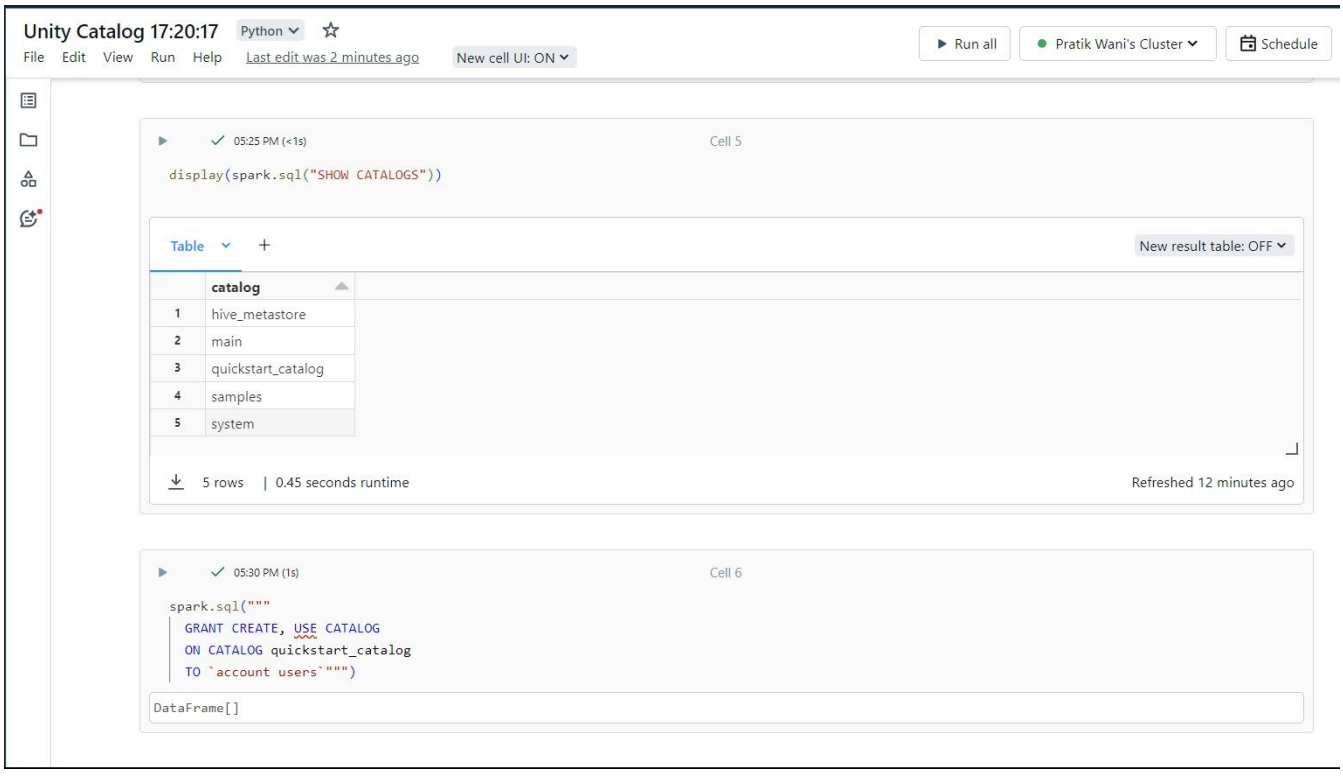
Also a metastore can be configured in ADLS Gen2 containes

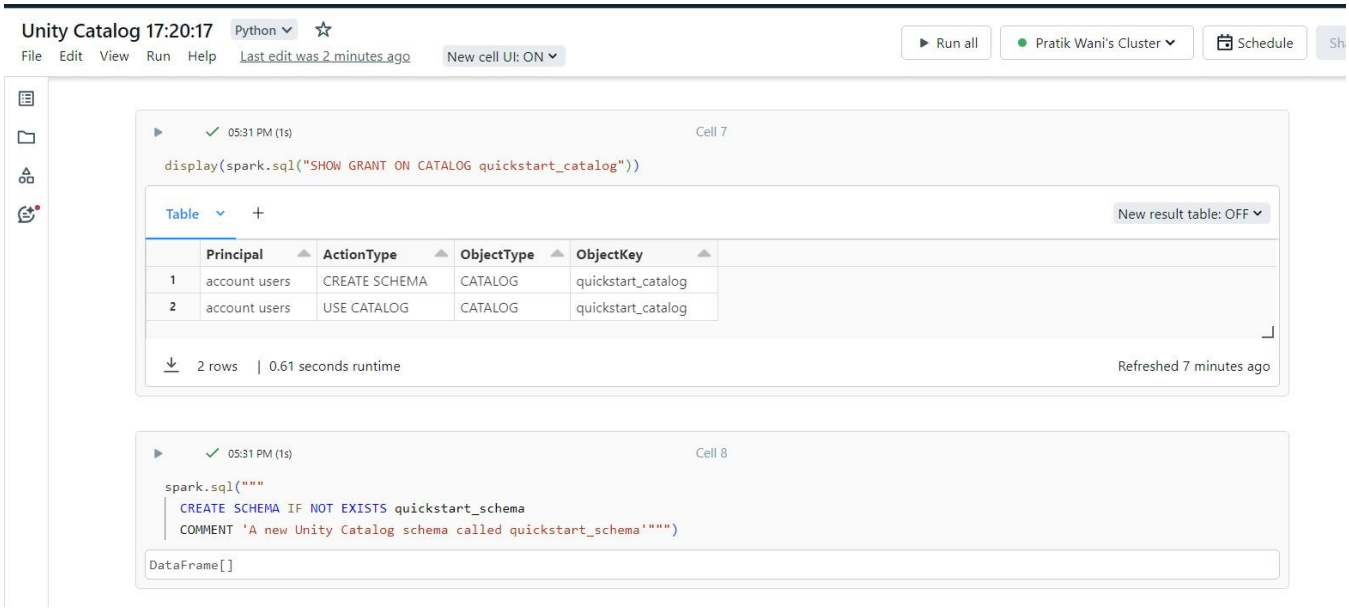
Externally managed data has title formate

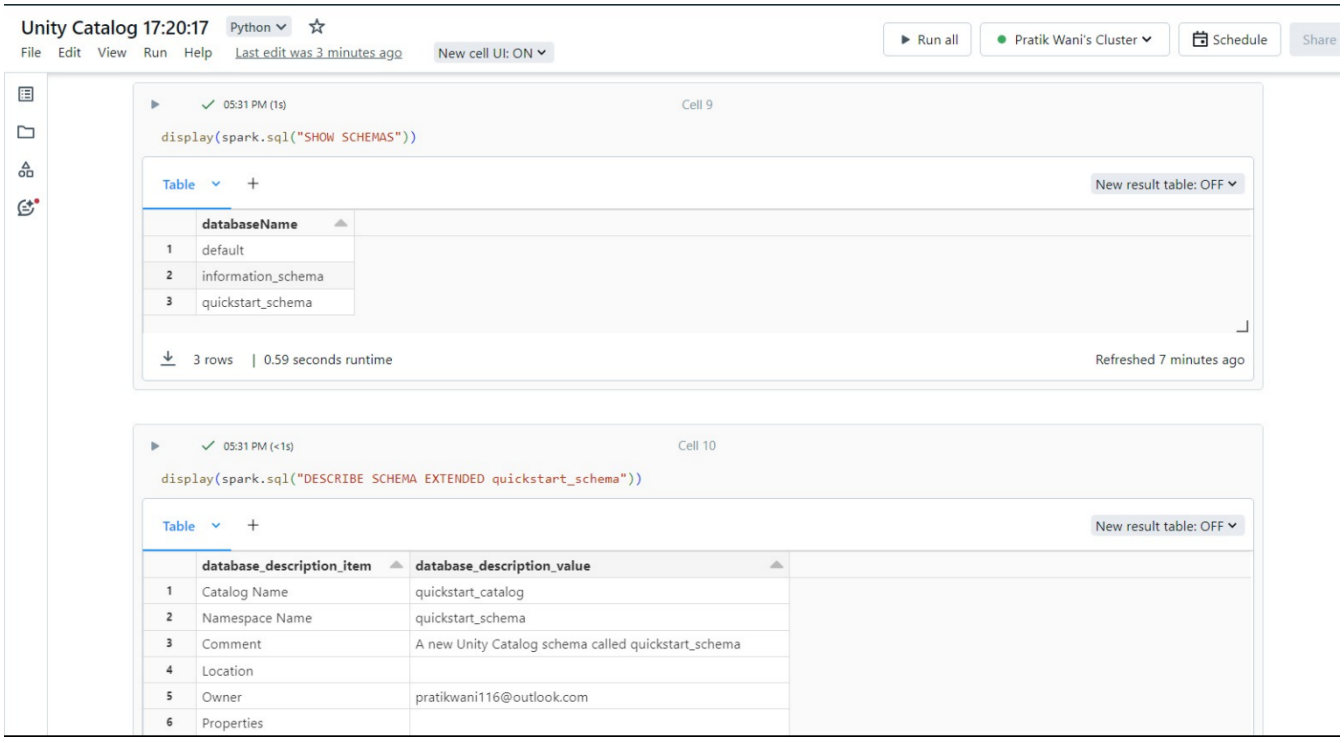
DELTA CSV AVRO PARAQUET
ORC TEXT

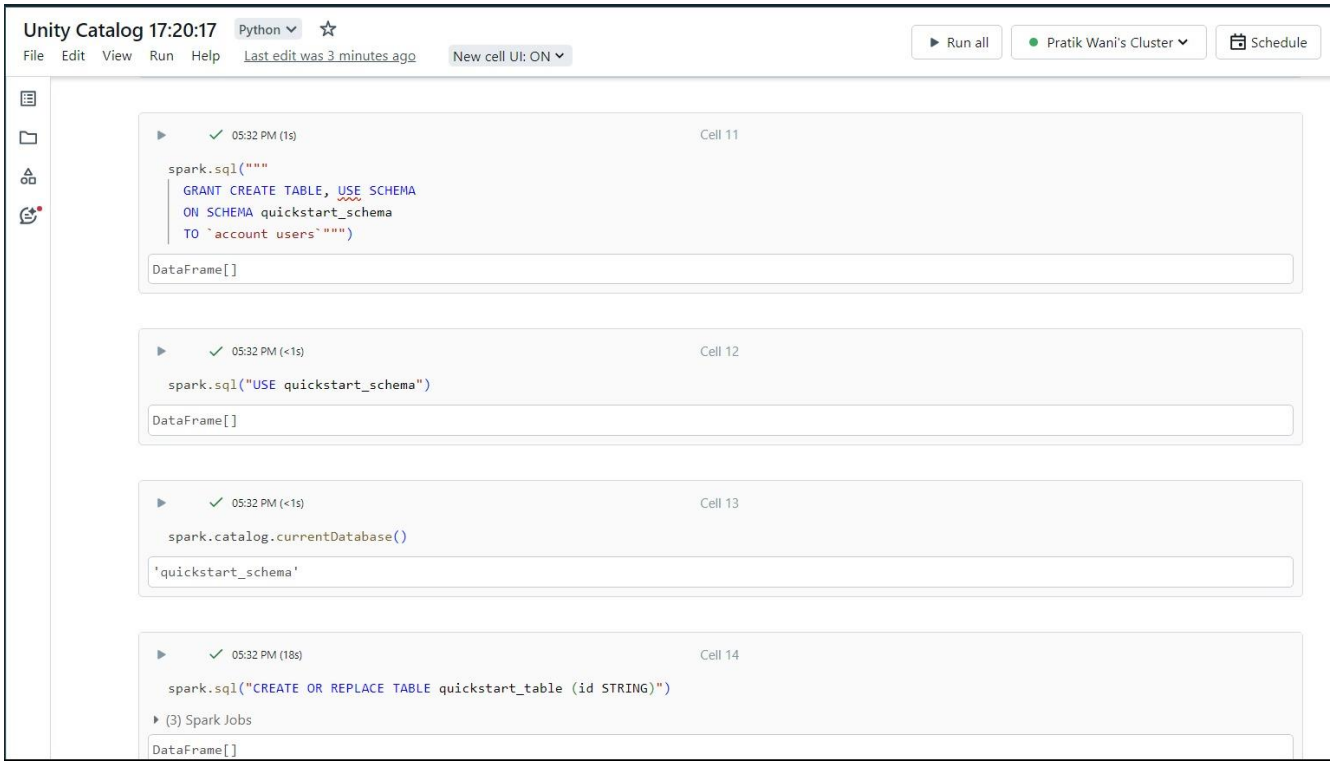
- Unity Catalog

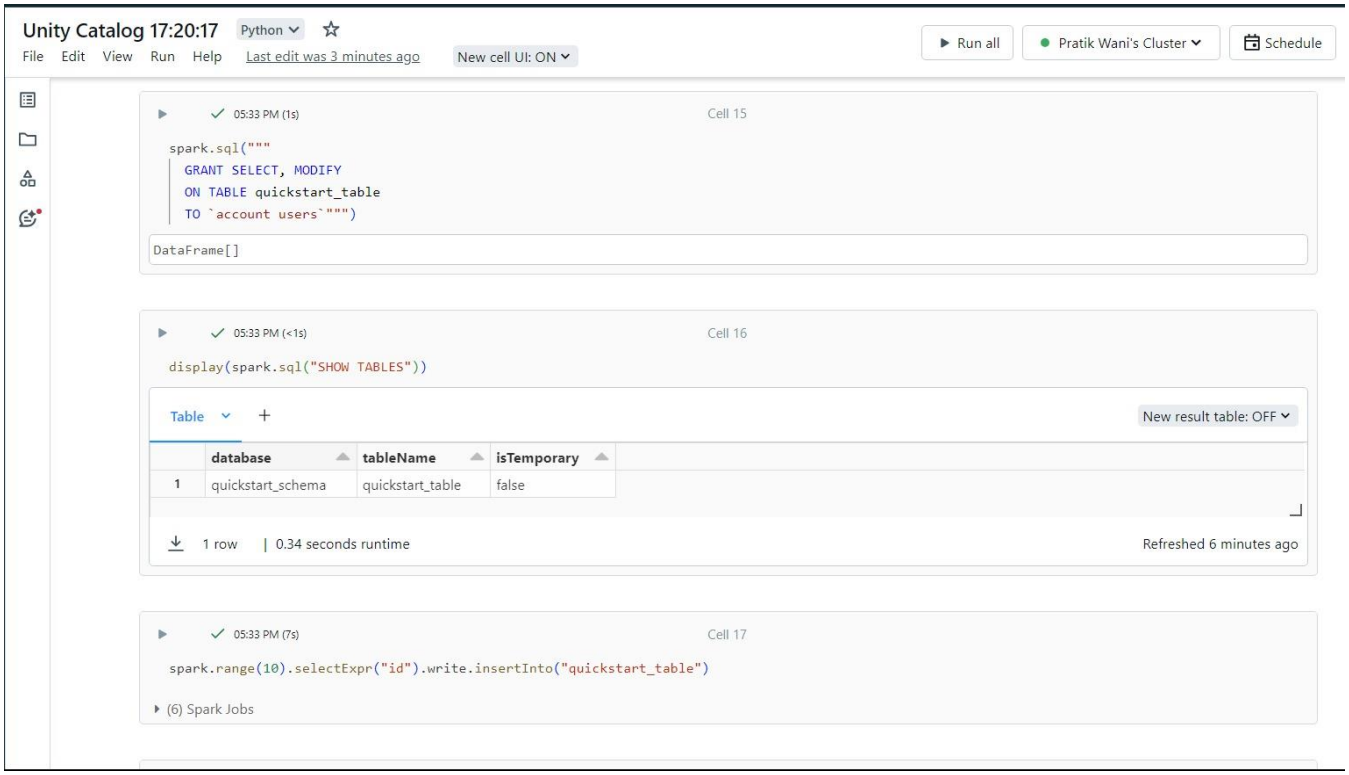


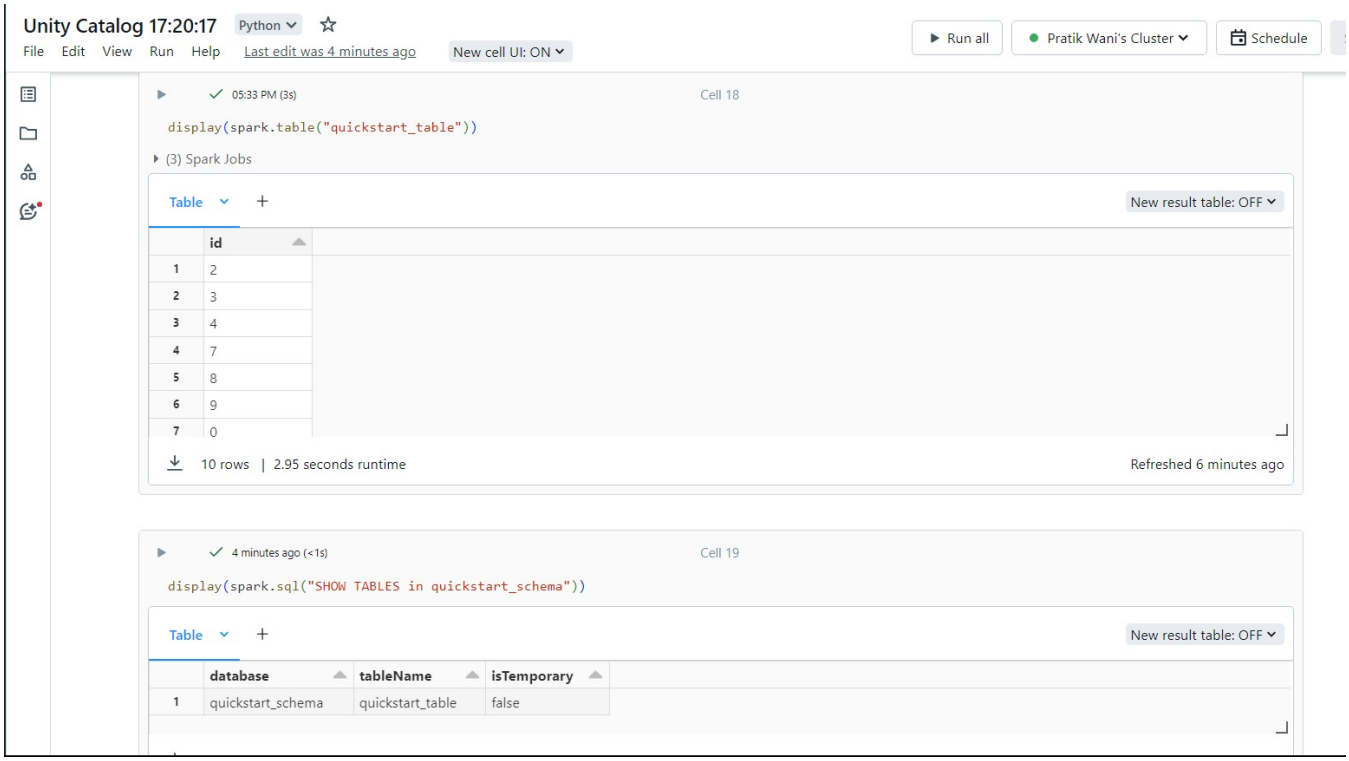


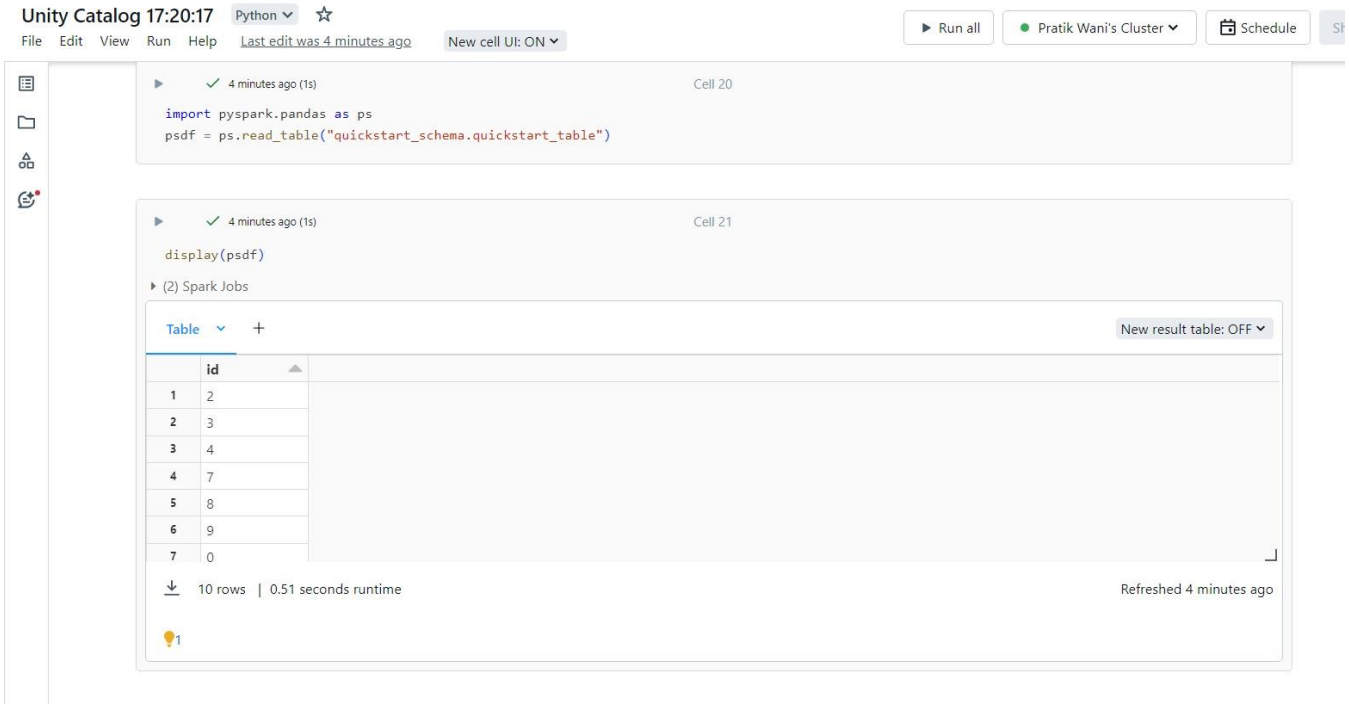

Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 2 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
05:31 PM (1s)
Cell 7
display(spark.sql("SHOW GRANT ON CATALOG quickstart_catalog"))
Table
New result table: OFF
Principal
ActionType
ObjectType
ObjectKey
1 account users CREATE SCHEMA CATALOG quickstart_catalog
2 account users USE CATALOG CATALOG quickstart_catalog
2 rows | 0.61 seconds runtime
Refreshed 7 minutes ago
05:31 PM (1s)
Cell 8
spark.sql("""
CREATE SCHEMA IF NOT EXISTS quickstart_schema
COMMENT 'A new Unity Catalog schema called quickstart_schema'""")
DataFrame[]

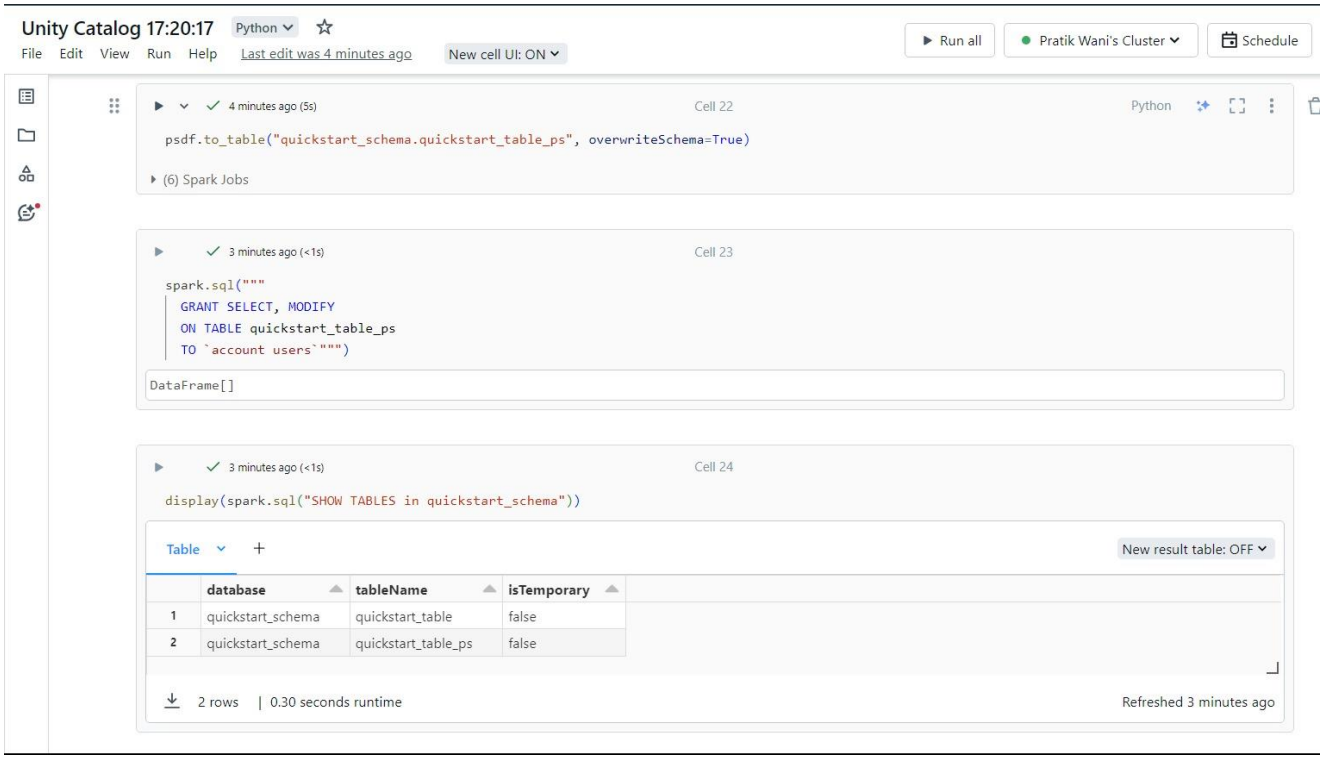

Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 3 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
Share
05:31 PM (1s)
Cell 9
display(spark.sql("SHOW SCHEMAS"))
Table
New result table: OFF
databaseName
1 default
2 information_schema
3 quickstart_schema
3 rows | 0.59 seconds runtime
Refreshed 7 minutes ago
05:31 PM (<1s)
Cell 10
display(spark.sql("DESCRIBE SCHEMA EXTENDED quickstart_schema"))
Table
New result table: OFF
database_description_item
database_description_value
1 Catalog Name quickstart_catalog
2 Namespace Name quickstart_schema
3 Comment A new Unity Catalog schema called quickstart_schema
4 Location
5 Owner pratikwani116@outlook.com
6 Properties

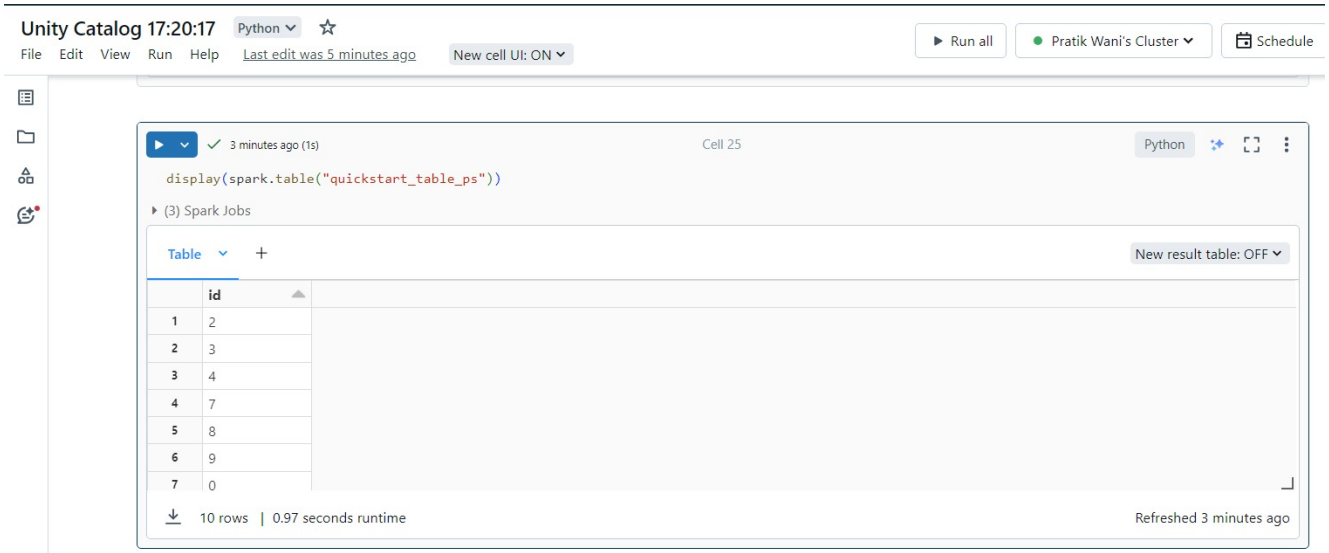

Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 3 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
05:32 PM (1s)
Cell 11
spark.sql("""
GRANT CREATE TABLE, USE SCHEMA
ON SCHEMA quickstart_schema
TO `account users`""")
DataFrame[]
05:32 PM (<1s)
Cell 12
spark.sql("USE quickstart_schema")
DataFrame[]
05:32 PM (<1s)
Cell 13
spark.catalog.currentDatabase()
'quickstart_schema'
05:32 PM (18s)
Cell 14
spark.sql("CREATE OR REPLACE TABLE quickstart_table (id STRING)")
(3) Spark Jobs
DataFrame[]


Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 3 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
05:33 PM (1s)
Cell 15
spark.sql("""
GRANT SELECT, MODIFY
ON TABLE quickstart_table
TO `account users`""")
DataFrame[]
05:33 PM (<1s)
Cell 16
display(spark.sql("SHOW TABLES"))
Table
New result table: OFF
database
tableName
isTemporary
1
quickstart_schema
quickstart_table
false
1 row | 0.34 seconds runtime
Refreshed 6 minutes ago
05:33 PM (7s)
Cell 17
spark.range(10).selectExpr("id").write.insertInto("quickstart_table")
(6) Spark Jobs

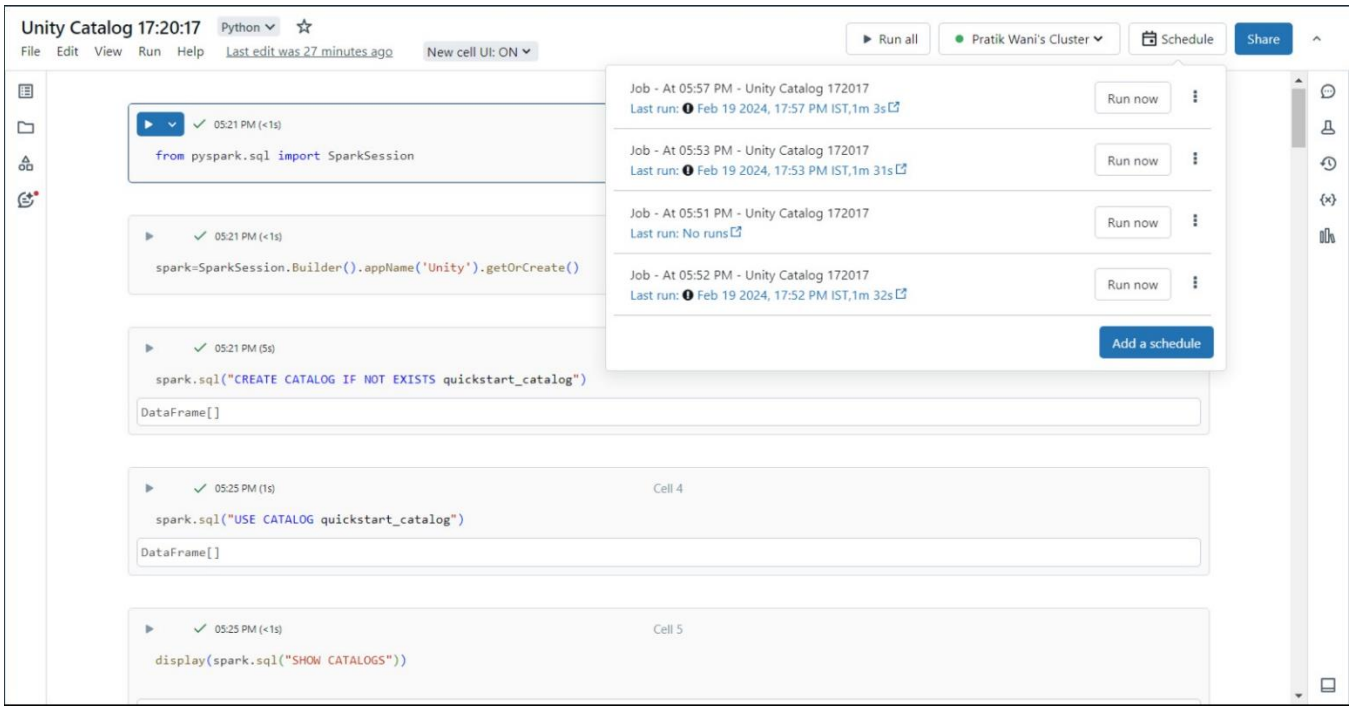

Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 4 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
05:33 PM (3s)
Cell 18
display(spark.table("quickstart_table"))
(3) Spark Jobs
Table
New result table: OFF
id
1 2
2 3
3 4
4 7
5 8
6 9
7 0
10 rows | 2.95 seconds runtime
Refreshed 6 minutes ago
4 minutes ago (<1s)
Cell 19
display(spark.sql("SHOW TABLES in quickstart_schema"))
Table
New result table: OFF
database
tableName
isTemporary
1 quickstart_schema quickstart_table false

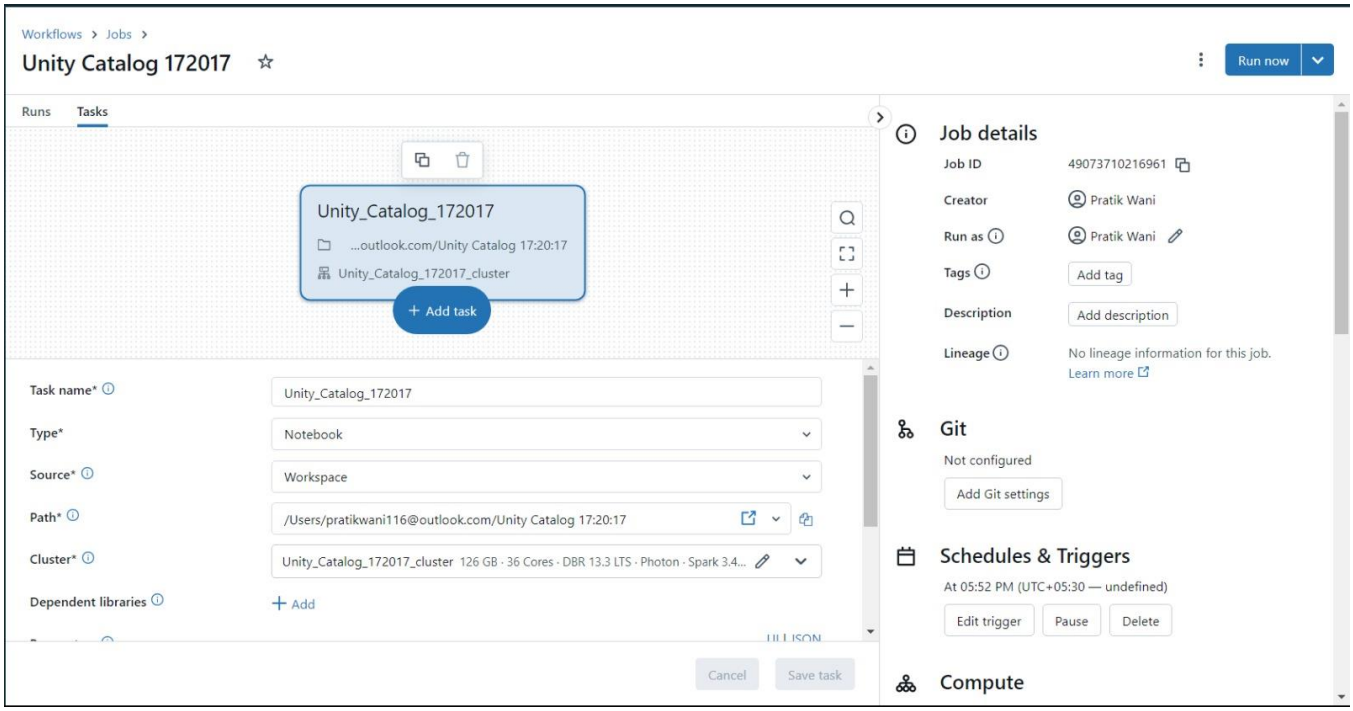

Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 4 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
4 minutes ago (1s)
Cell 20
import pyspark.pandas as ps
psdf = ps.read_table("quickstart_schema.quickstart_table")
4 minutes ago (1s)
Cell 21
display(psdf)
(2) Spark Jobs
Table
New result table: OFF
id
1 2
2 3
3 4
4 7
5 8
6 9
7 0
10 rows | 0.51 seconds runtime
Refreshed 4 minutes ago

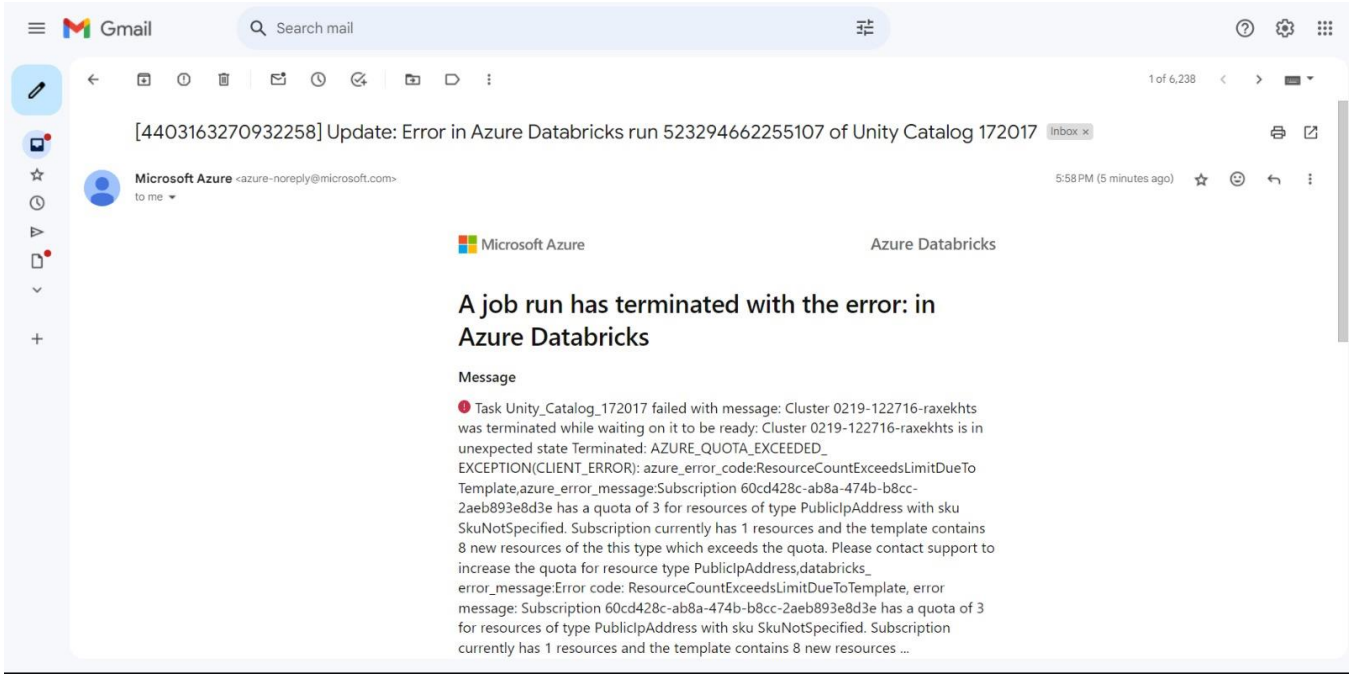

Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 4 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
4 minutes ago (5s)
Cell 22
Python
psdf.to_table("quickstart_schema.quickstart_table_ps", overwriteSchema=True)
(6) Spark Jobs
3 minutes ago (<1s)
Cell 23
spark.sql("""
GRANT SELECT, MODIFY
ON TABLE quickstart_table_ps
TO `account users`""")
DataFrame[]
3 minutes ago (<1s)
Cell 24
display(spark.sql("SHOW TABLES in quickstart_schema"))
Table
New result table: OFF
database
tableName
isTemporary
1 quickstart_schema quickstart_table false
2 quickstart_schema quickstart_table_ps false
2 rows | 0.30 seconds runtime
Refreshed 3 minutes ago


Unity Catalog 17:20:17
Python
File Edit View Run Help
Last edit was 5 minutes ago
New cell UI: ON
Run all
Pratik Wani's Cluster
Schedule
3 minutes ago (1s)
Cell 25
Python
display(spark.table("quickstart_table_ps"))
(3) Spark Jobs
Table
New result table: OFF
id
1 2
2 3
3 4
4 7
5 8
6 9
7 0
10 rows | 0.97 seconds runtime
Refreshed 3 minutes ago

- Job Scheduling






Gmail
Search mail
1 of 6,238
[4403163270932258] Update: Error in Azure Databricks run 523294662255107 of Unity Catalog 172017
Inbox ×
Microsoft Azure <azure-noreply@microsoft.com>
to me
5:58 PM (5 minutes ago)
Microsoft Azure
Azure Databricks
A job run has terminated with the error: in Azure Databricks
Message
Task Unity_Catalog_172017 failed with message: Cluster 0219-122716-raxekhts was terminated while waiting on it to be ready: Cluster 0219-122716-raxekhts is in unexpected state Terminated: AZURE_QUOTA_EXCEEDED_EXCEPTION(CLIENT_ERROR): azure_error_code:ResourceCountExceedsLimitDueToTemplate,azure_error_message:Subscription 60cd428c-ab8a-474b-b8cc-2aeb893e8d3e has a quota of 3 for resources of type PublicIpAddress with sku SkuNotSpecified. Subscription currently has 1 resources and the template contains 8 new resources of the this type which exceeds the quota. Please contact support to increase the quota for resource type PublicIpAddress,databricks_error_message:Error code: ResourceCountExceedsLimitDueToTemplate, error message: Subscription 60cd428c-ab8a-474b-b8cc-2aeb893e8d3e has a quota of 3 for resources of type PublicIpAddress with sku SkuNotSpecified. Subscription currently has 1 resources and the template contains 8 new resources ...