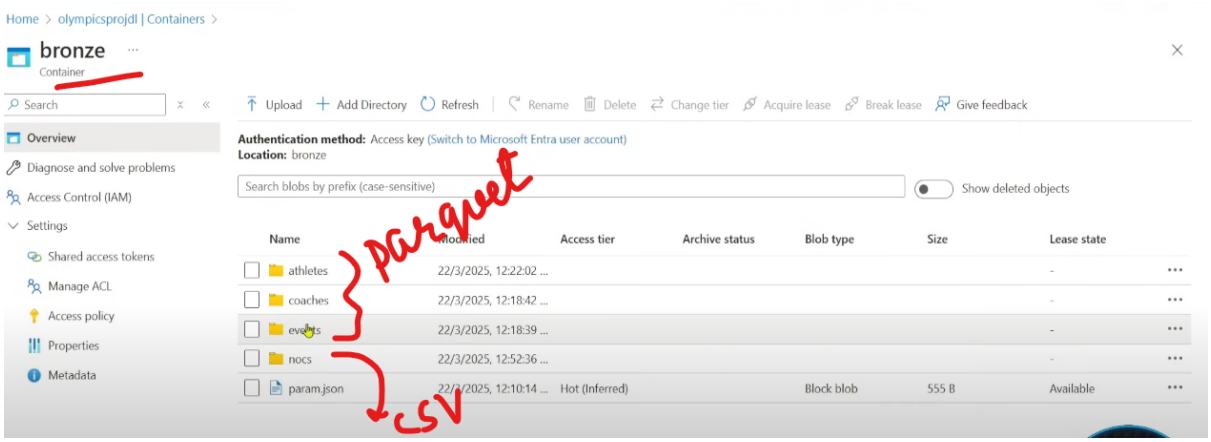
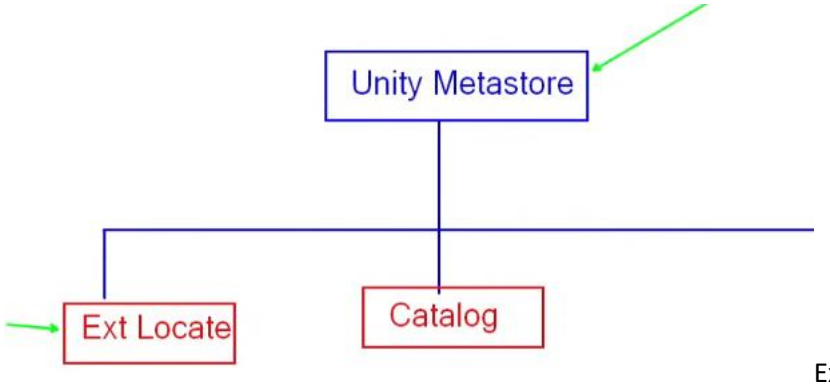


Silver Notebook

Reading NOCS Data

It was able to read the data using External Location for bronze container, but how DB knows that we need to use Bronze External Location



Ext Loc is connected to UM, so it will try to hunt the ext location for that container & use it. And if no ext loc it will throw an error.

Dropping the column

Table +

	code	country_long	tag	note
1	AFG	Afghanistan	afghanistan	P
2	AHO	Netherlands Antilles	netherlands-antilles	H
3	AIN	AIN	ain	P
4	ALB	Albania	albania	P

Waiting

```
df = df.withColumn('tag',split(col('tag'),'-')[0])
df.display()
```

Just now (5s) 10

```
df.write.format("delta")\
    .mode("append")\
    .option("path","abfss://silver@olympicsprojdl.dfs.core.windows.net/nocs")\
    .save()
```

See performance (1)

Just now (2s) 10

```
df.write.format("delta")\
    .mode("append")\
    .option("path","abfss://silver@olympicsprojdl.dfs.core.windows.net/nocs")\
    .saveAsTable("olympics.silver.nocs")
```

width layout See performance (1)

creating ext table

Data will be in my container Data Lake but table metadata in metastore

adb-4181826646446634.14.azuredatabricks.net/explore/data/olympics/silver/nocs?o=4181826646446634&activeTab=details

Microsoft Azure databricks Search data, notebooks, recents, and more... CTRL + P

Catalog

Serverless Starter Warehouse Serverless

Type to search...

- My organization
  - system
  - main
  - olympics
    - default
    - information_schema
    - silver
      - nocs
      - nocs_man
- Delta Shares Received
  - samples
- Legacy
  - hive_metastore

Catalog Explorer › olympics › silver ›

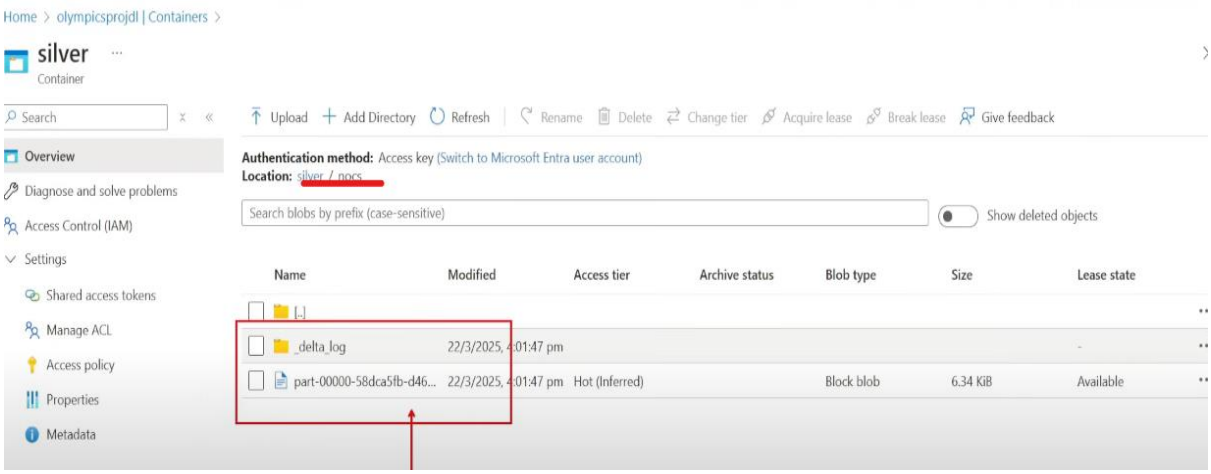
nocs

Open in a dashbo

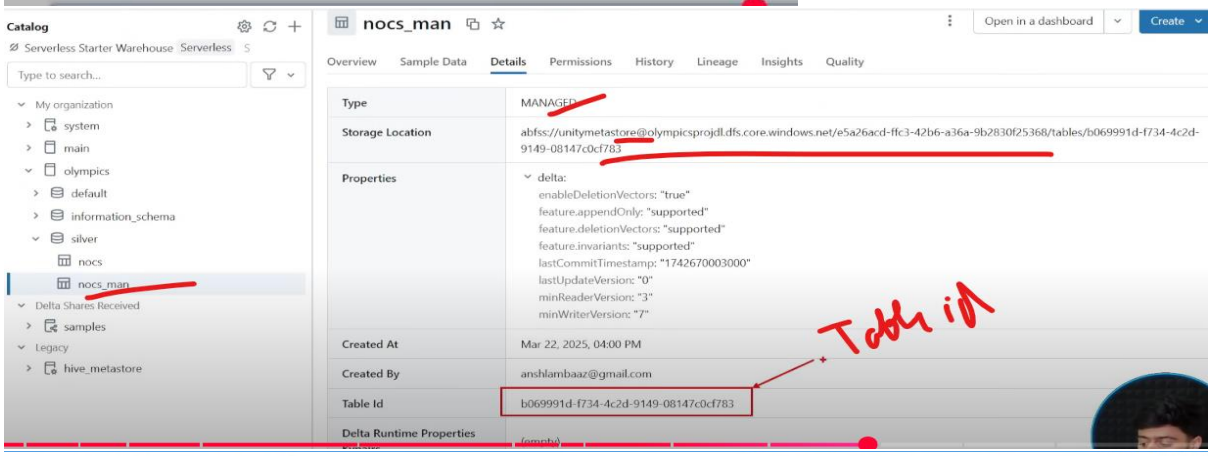
Overview Sample Data Details Permissions History Lineage Insights Quality

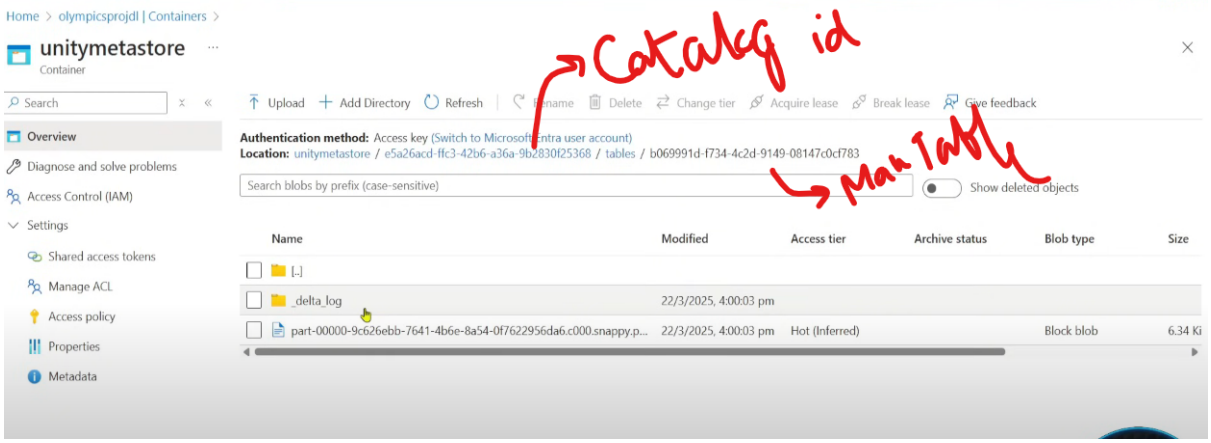
Type	EXTERNAL
Storage Location	abfss://silver@olympicsprojdl.dfs.core.windows.net/nocs
Properties	delta.enableDeletionVectors: "true" delta.feature.appendOnly: "supported" delta.feature.deletionVectors: "supported" delta.feature.invariants: "supported" delta.lastCommitTimestamp: "1742670108000" delta.lastUpdateVersion: "0" delta.minReaderVersion: "3" delta.minWriterVersion: "7"
Created At	Mar 22, 2025, 04:01 PM
Created By	anshlambaaz@gmail.com
Updated At	Mar 22, 2025, 04:01 PM
Updated By	anshlambaaz@gmail.com
Table Id	7d70cece-1d95-41b8-852f-2a1584a006b2

Table Id

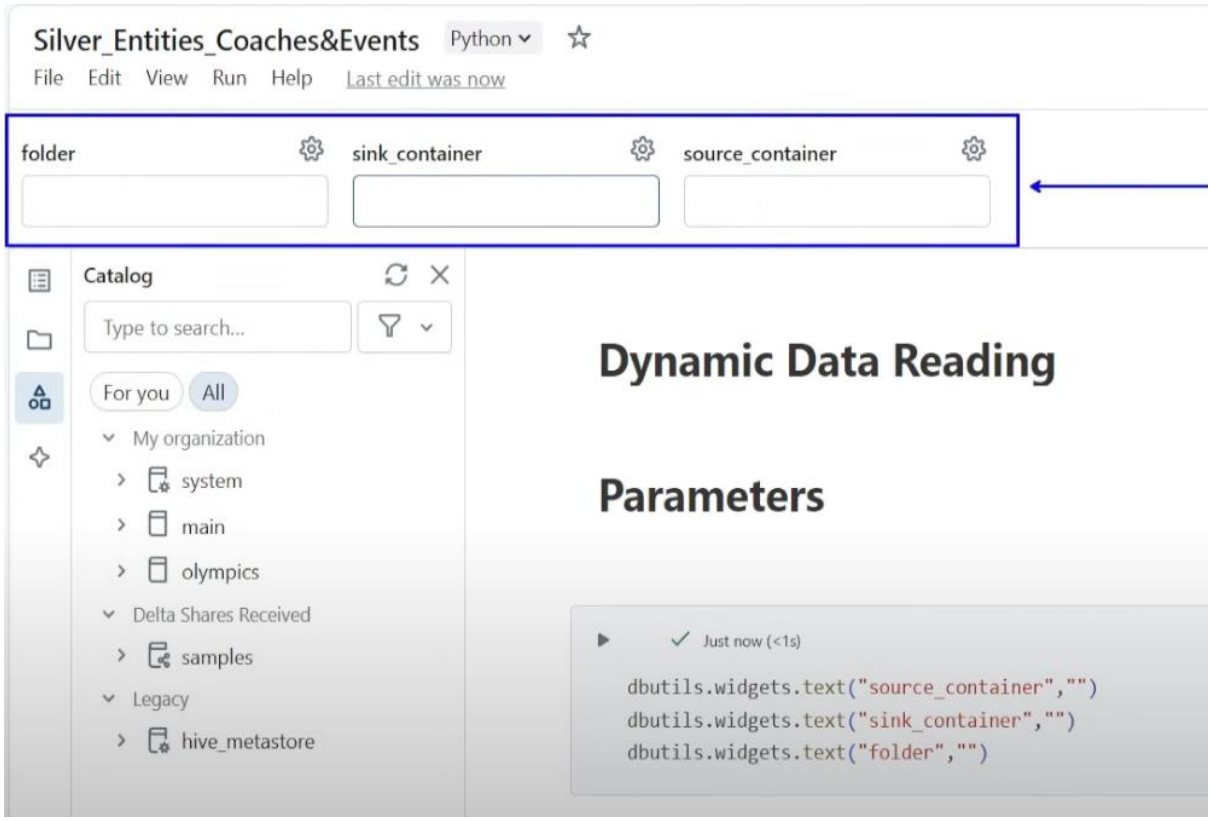


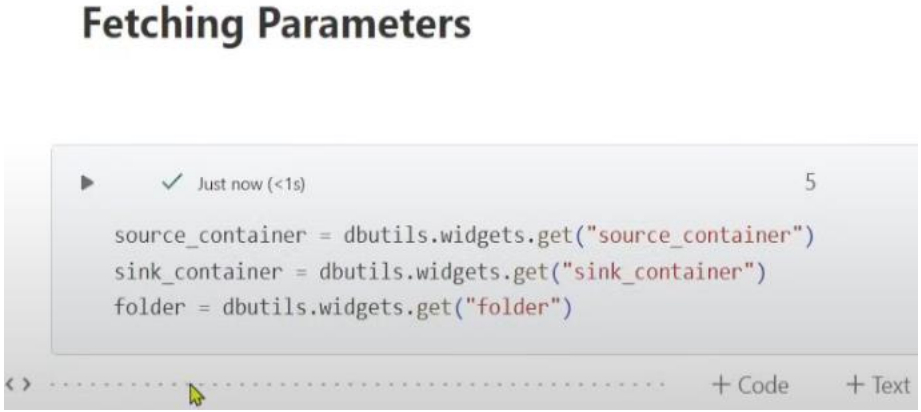
Extra→ Lets create Managed Table

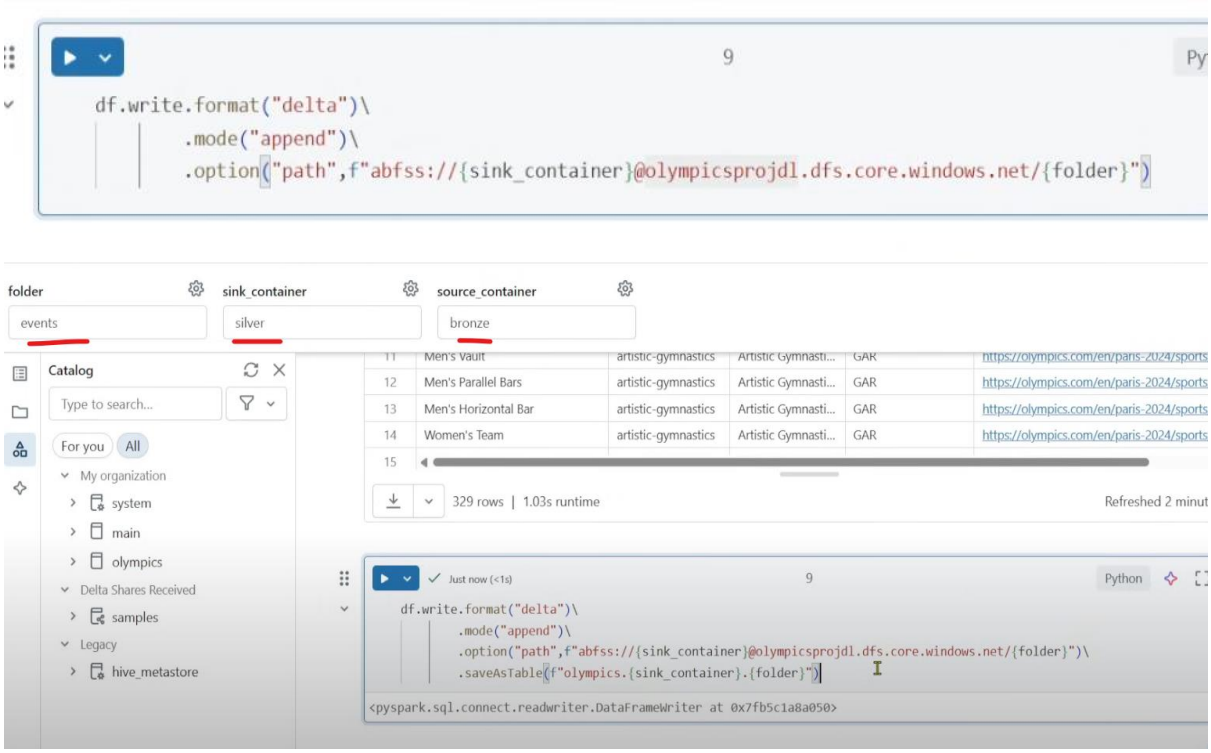




For Coaches & Events, lets read the data dynamically i.e. parameterized

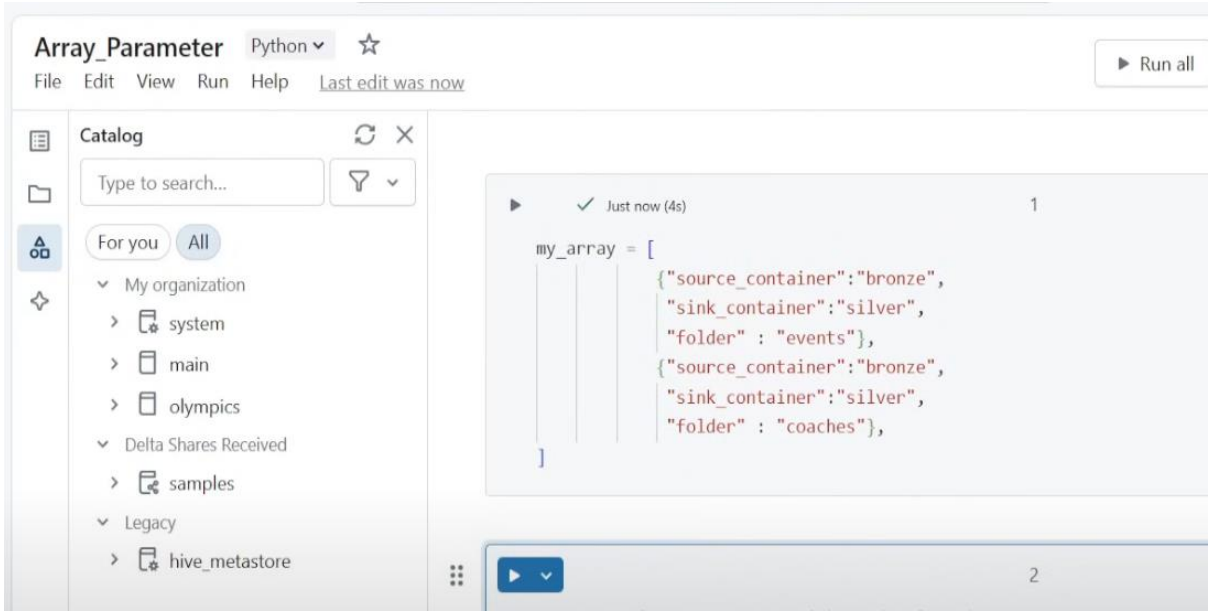




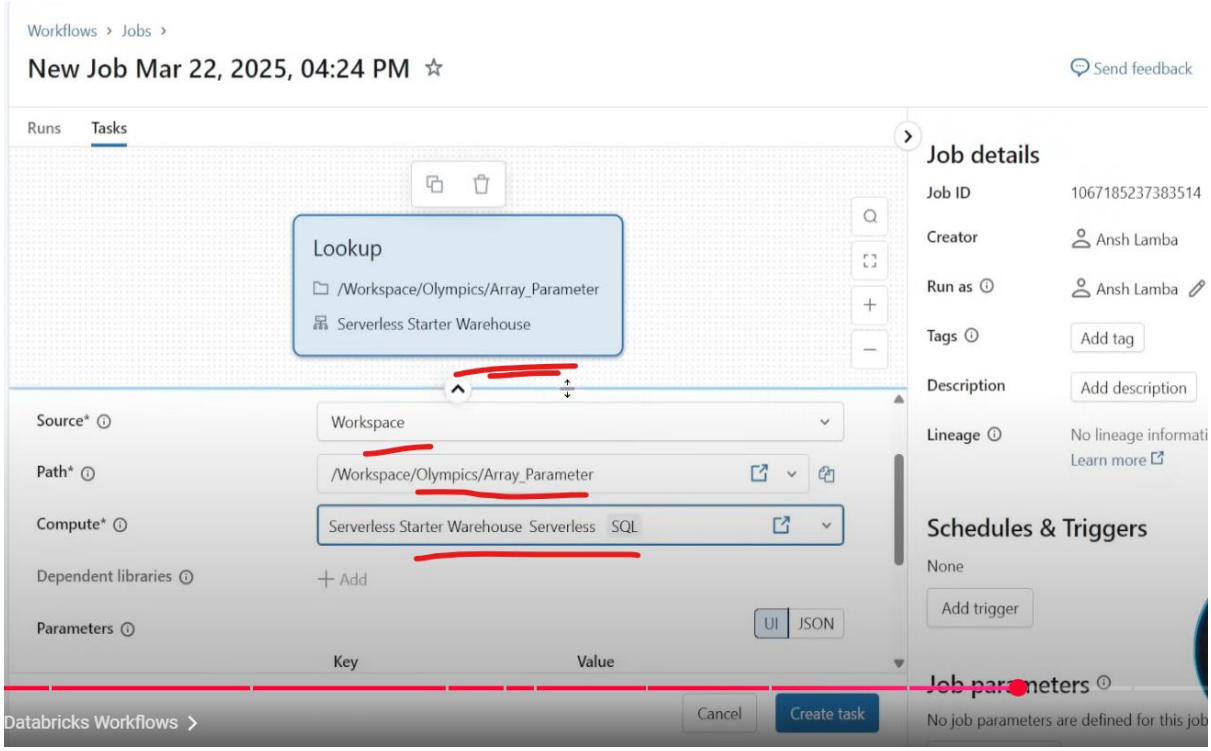


Now we want to pass the parameters from Workflows

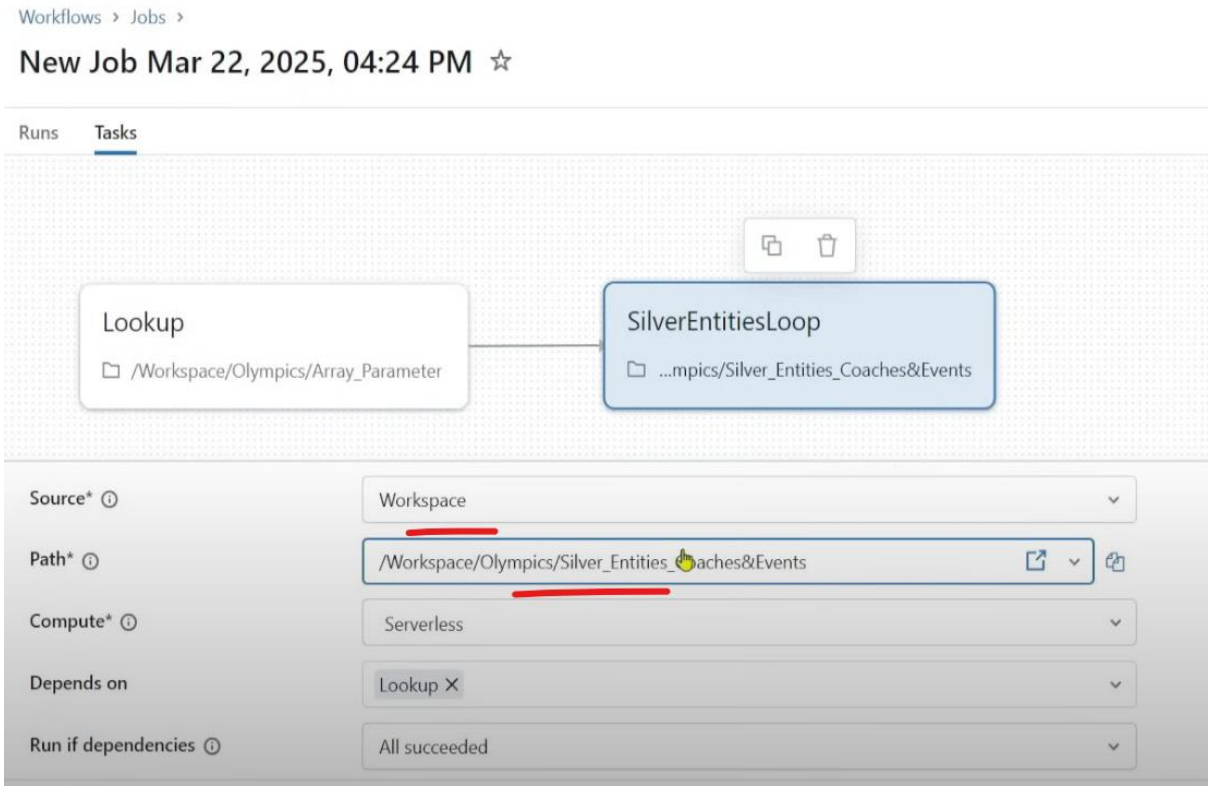
Lets create another notebook to hold the values of array

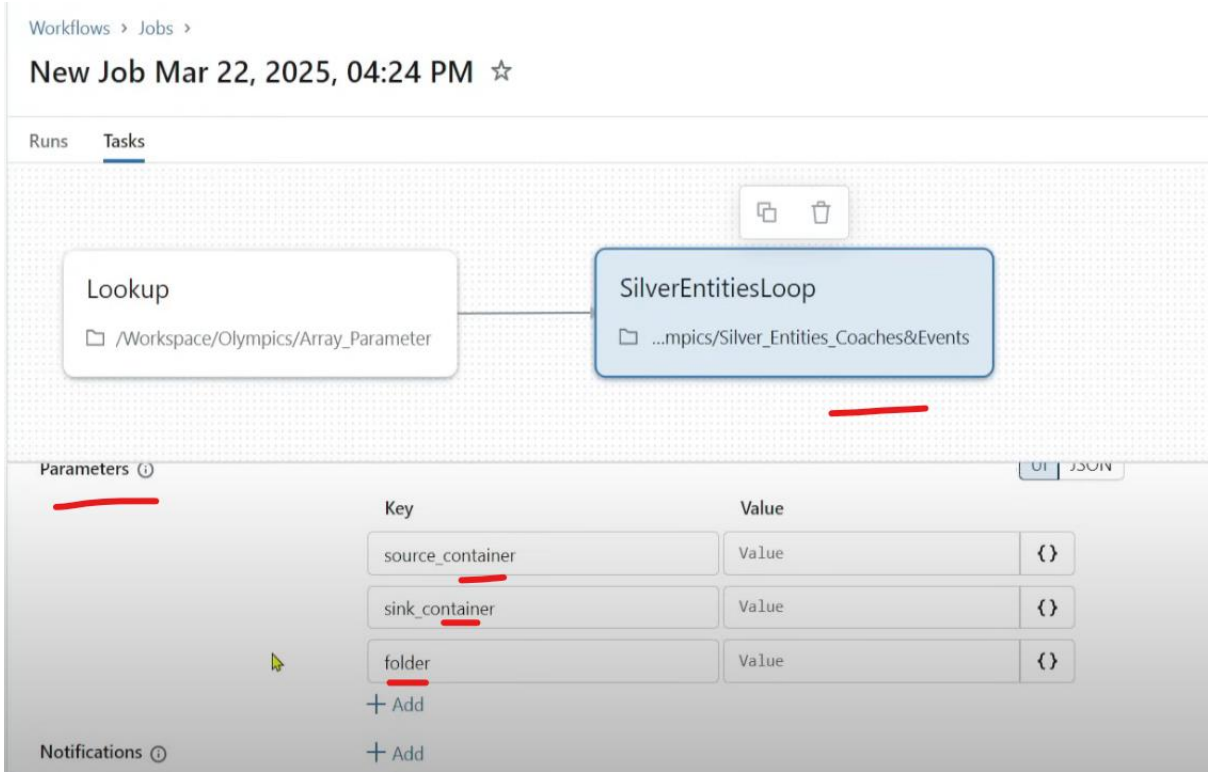


Now go to Workflows → Create Jobs →

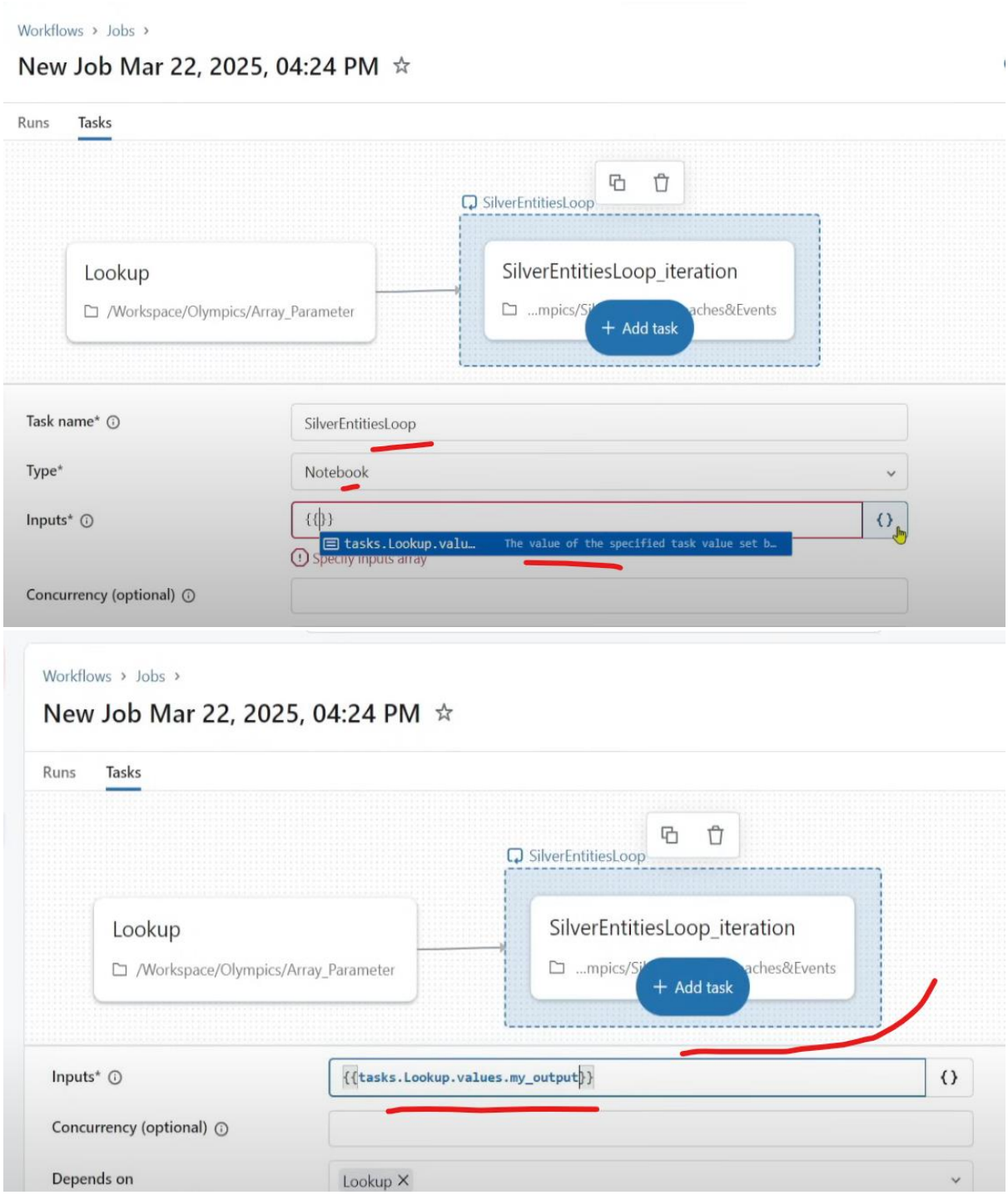


Create Task → Add new task

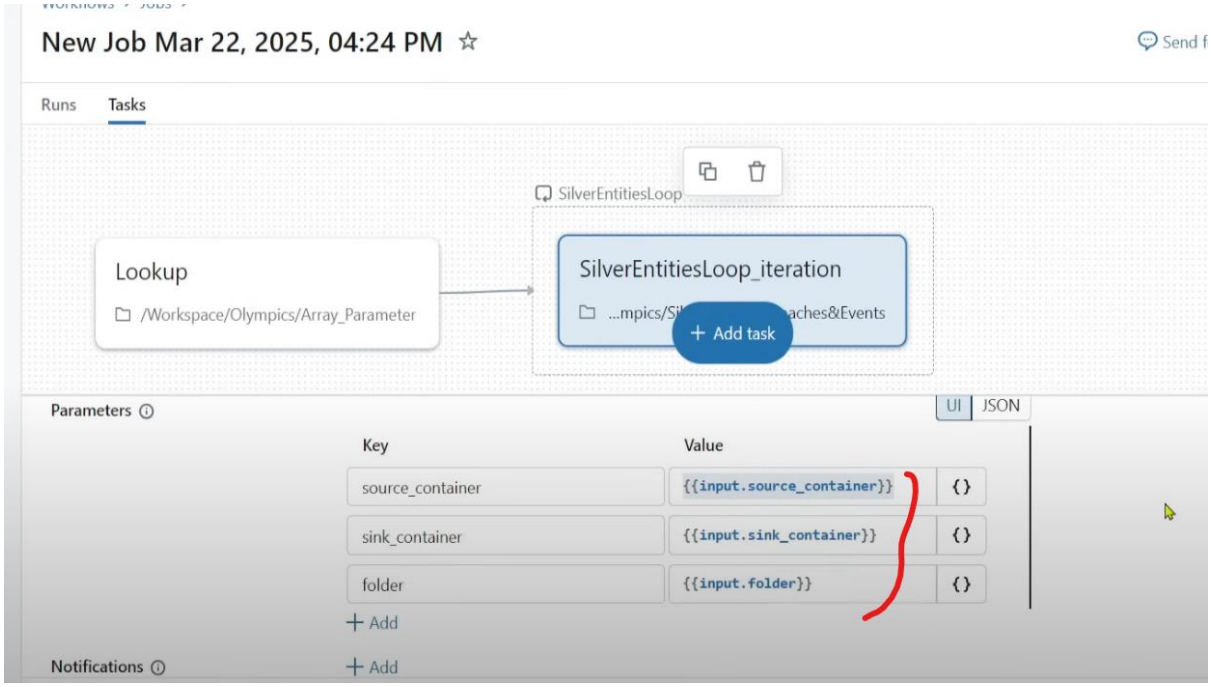




Value is coming from lookup in the form of array, but in value of parameters we can pass a single item only, so we have to loop through the output of Lookup

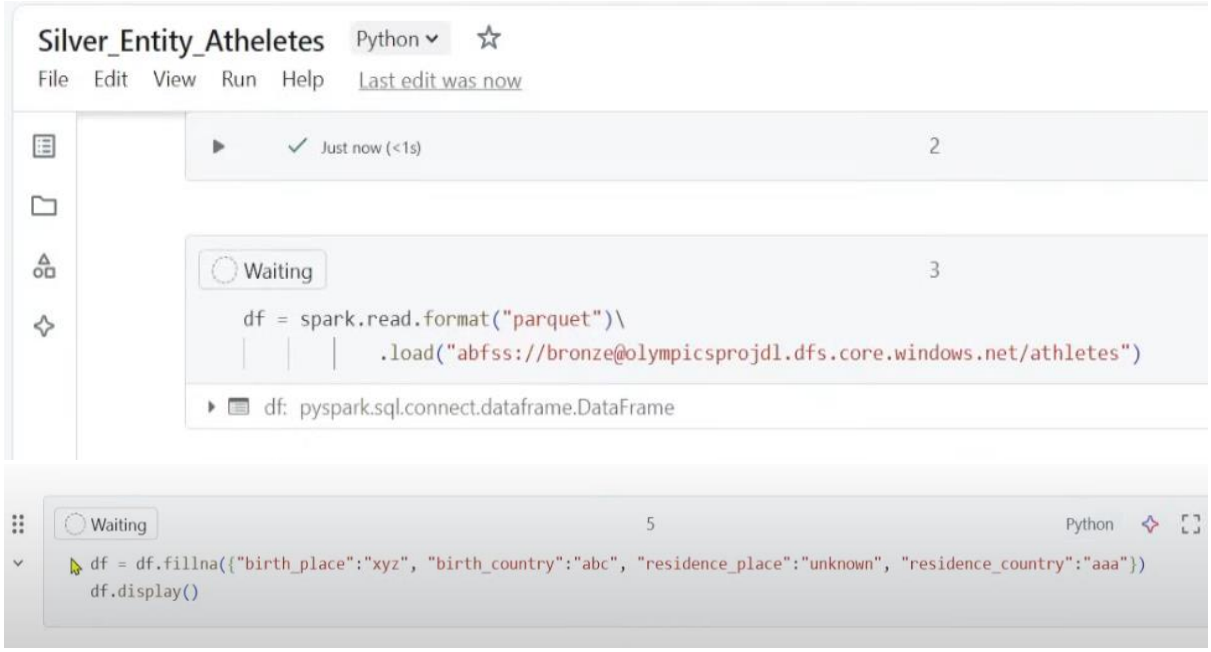


Now go to the Notebook activity inside the Loop



Save and Run the Workflow

Now lets work on the Athletes file (bronze to silver with some transformation)



```
Just now (1s)                                                    6
df_filtered = df.filter((col('current')==True) & (col('name').isin('GALSTYAN Slavik','HARUTYUNYAN Arsen','SEHEN Sajjad')))
df_filtered.display()
> See performance (1)
df_filtered:  pyspark.sql.connect.dataframe.DataFrame = [code: string, current: string ... 34 more fields]
```

Table +

	code	current	name	name_short	name_tv	gender	function	country
1	1532874	True	GALSTYAN Slavik	GALSTYAN S	Slavik GALSTYAN	Male	Athlete	ARM
2	1532944	True	HARUTYUNYAN Arsen	HARUTYUNYAN A	Arsen HARUTYUNYAN	Male	Athlete	ARM
3	1533245	True	SEHEN Sajjad	SEHEN S	Sajjad SEHEN	Male	Athlete	IRQ

```
Interrupt  00:01                                                 7
df = df.withColumn('height',col('height').cast(FloatType()))\
       .withColumn('weight',col('weight').cast(FloatType()))

df.display()
```

```
Waiting                                                          8
1  df_sorted = df.sort('height','weight',ascending=[0,1]).filter(col('weight')>0)
2  df_sorted.display()
```

```
Waiting                                                          9
df_sorted = df_sorted.withColumn('nationality',regexp_replace('nationality','United States', 'US'))
df_sorted.display()
```

To find duplicate on a column

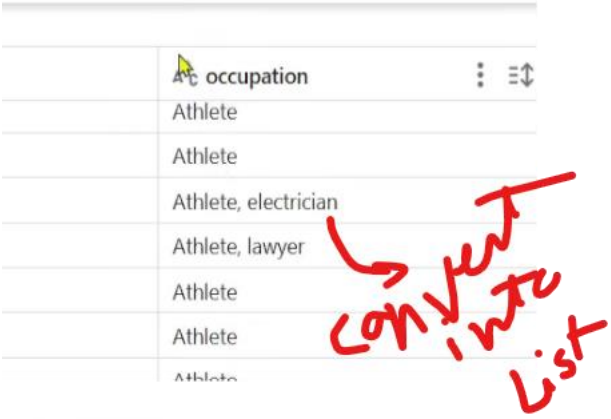
```
Just now (1s)                                                    1
df.groupBy('code').agg(count('code')).display()
> See performance (1)
```

Table +

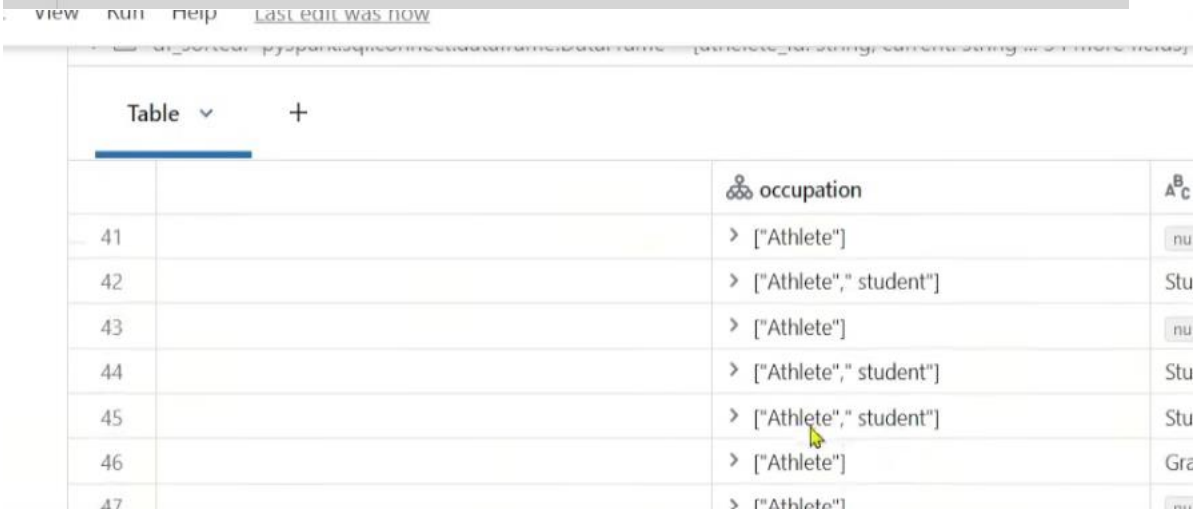
	code	count(code)
1	1536971	1
2	1562800	1
3	1888708	1
4	1888980	1
5	1901740	1

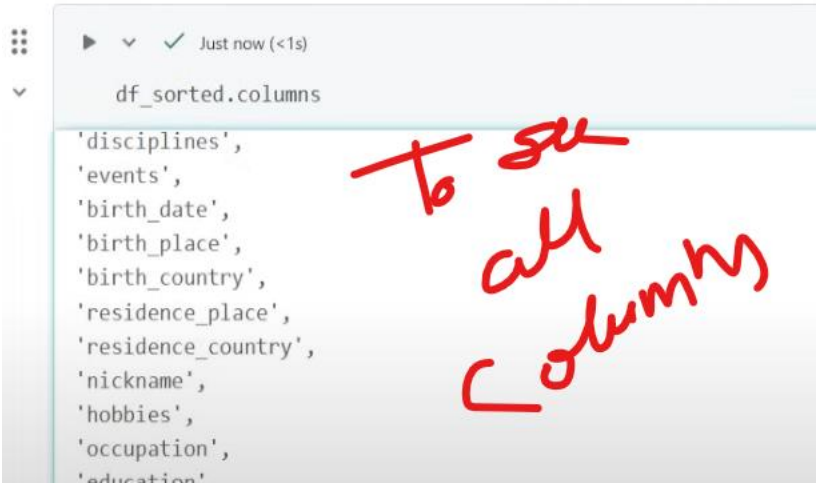

Just now (1s)
10
df.groupBy('code').agg(count('code').alias('total_count')).filter(col('total_count')>1).display()
See performance (1)
Table
code
total_count


Waiting
11
df_sorted = df_sorted.withColumnRenamed("code","athelete_id")
df_sorted.display()

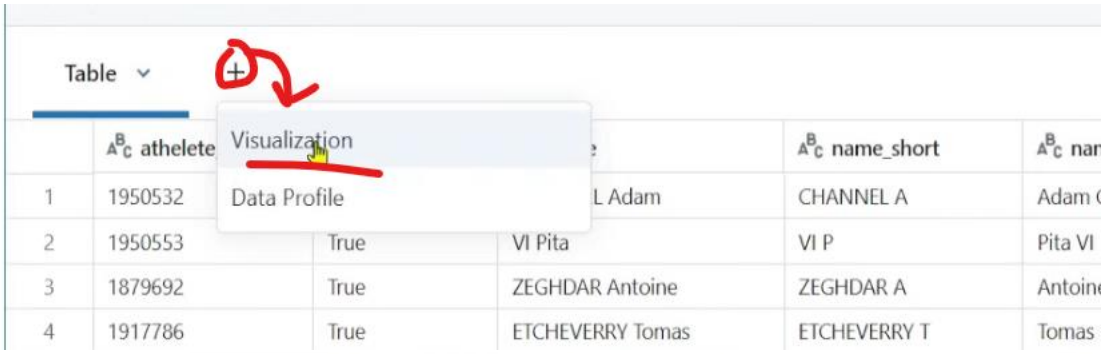

occupation
Athlete
Athlete
Athlete, electrician
Athlete, lawyer
Athlete
Athlete
convert into List


Just now (1s)
12
df_sorted = df_sorted.withColumn('occupation',split('occupation',','))
df_sorted.display()

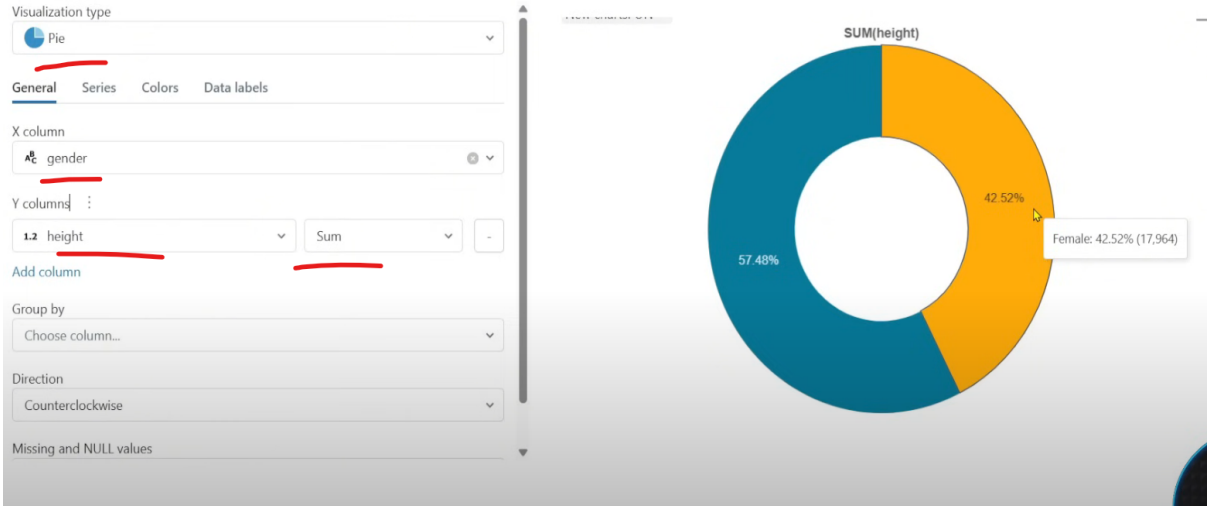

View
Run
Help
Last edit was now
Table
occupation
41
["Athlete"]
42
["Athlete"," student"]
43
["Athlete"]
44
["Athlete"," student"]
45
["Athlete"," student"]
46
["Athlete"]

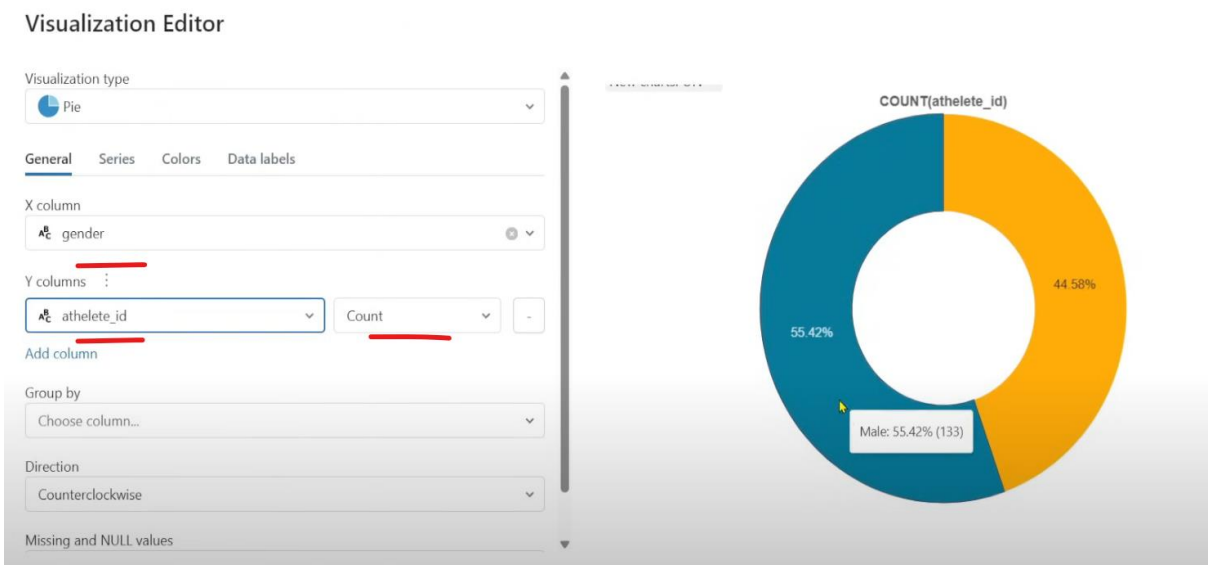


Visualization in ADB

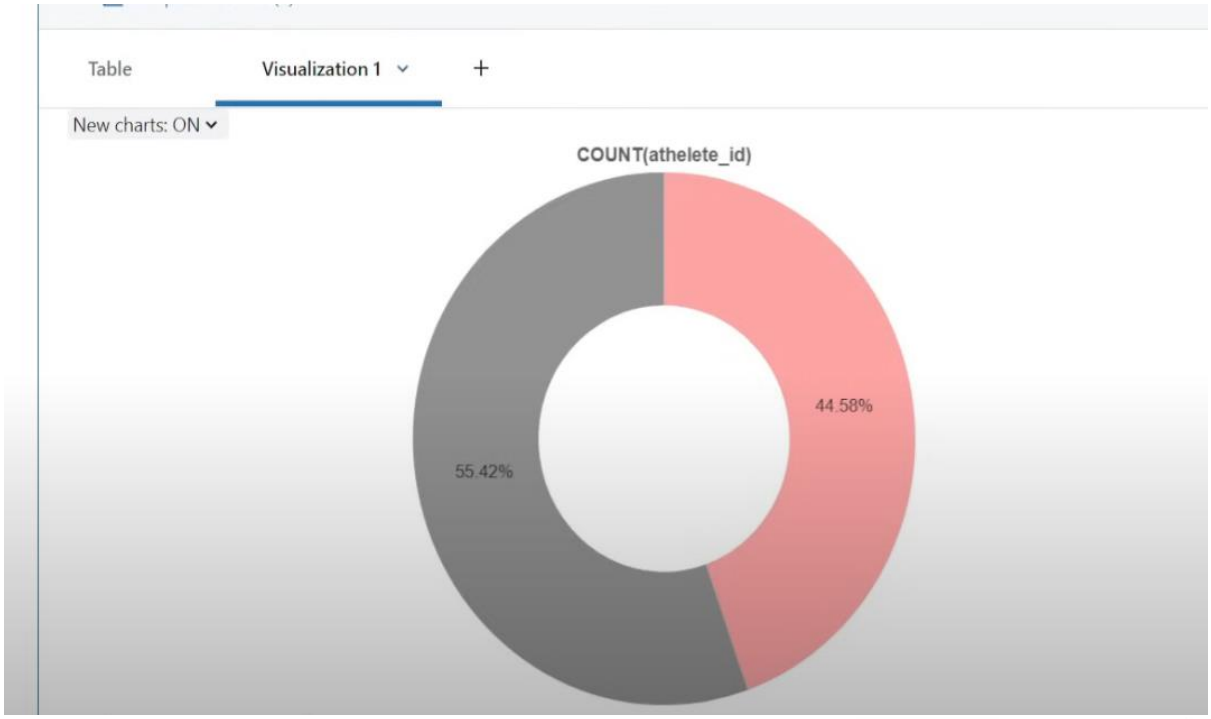


Visualization Editor





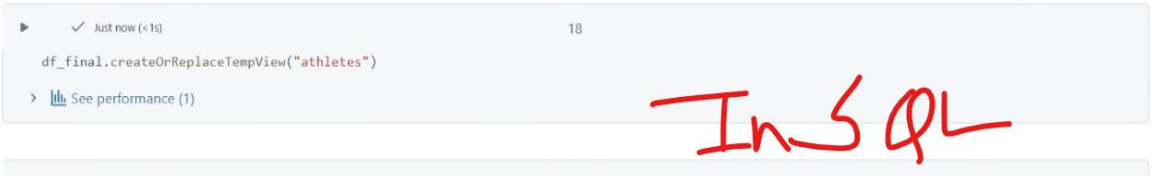
Save it to use it as a Dashboard in ADB only



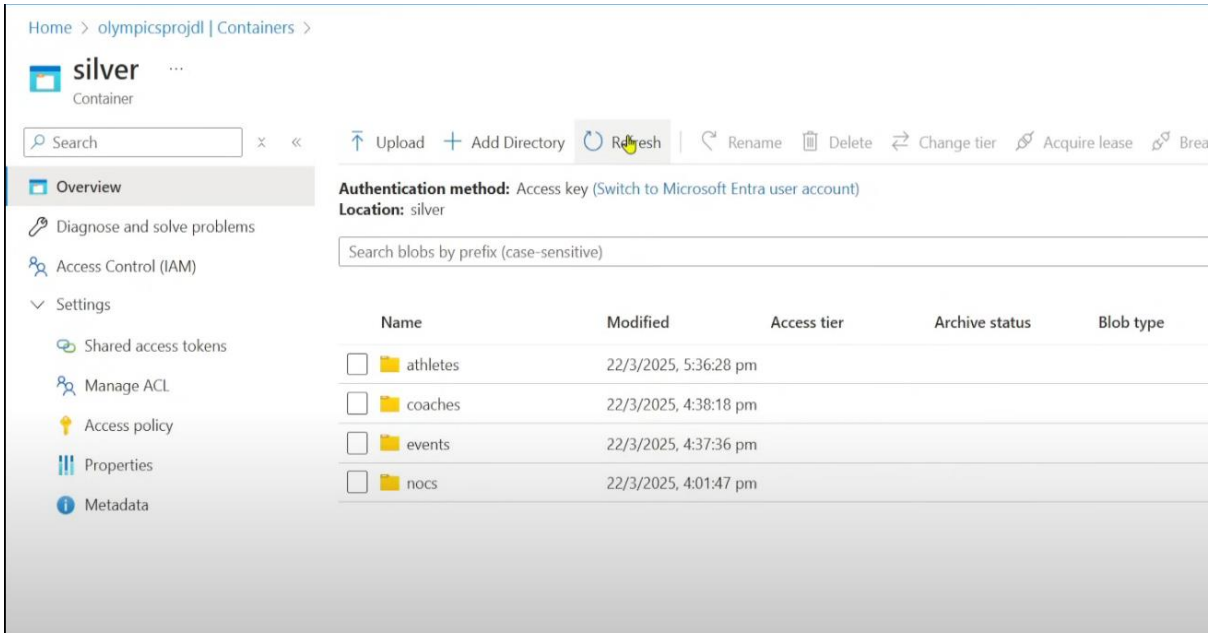
	ry	country_long	nationality	nationality_long	nationality_code	1.2 height	1.2 weight	1.2 cum_weight
1		Argentina	Argentina	Argentina	ARG	170	76	76
2		Argentina	Argentina	Argentina	ARG	172	72	148
3		Argentina	Argentina	Argentina	ARG	180	95	243
4		Argentina	Argentina	Argentina	ARG	181	82	325
5		Argentina	Argentina	Argentina	ARG	182	90	415

Table +

	ry	country_long	nationality	nationality_long	nationality_code	height	weight	cum_weight
9		Argentina	Argentina	Argentina	ARG	187	93	1081
10		Argentina	Argentina	Argentina	ARG	191	96	1081
11		Argentina	Argentina	Argentina	ARG	192	98	1081
12		Argentina	Argentina	Argentina	ARG	195	105	1081
13		Australia	Australia	Australia	AUS	164	62	1358
14		Australia	Australia	Australia	AUS	167	56	1358
15		Australia	Australia	Australia	AUS	170	72	1358
16		Australia	Australia	Australia	AUS	172	74	1358
17		Australia	Australia	Australia	AUS	173	66	1358
18		Australia	Australia	Australia	AUS	173	78	1358
19		Australia	Australia	Australia	AUS	174	70	1358
20		Australia	Australia	Australia	AUS	175	84	1358

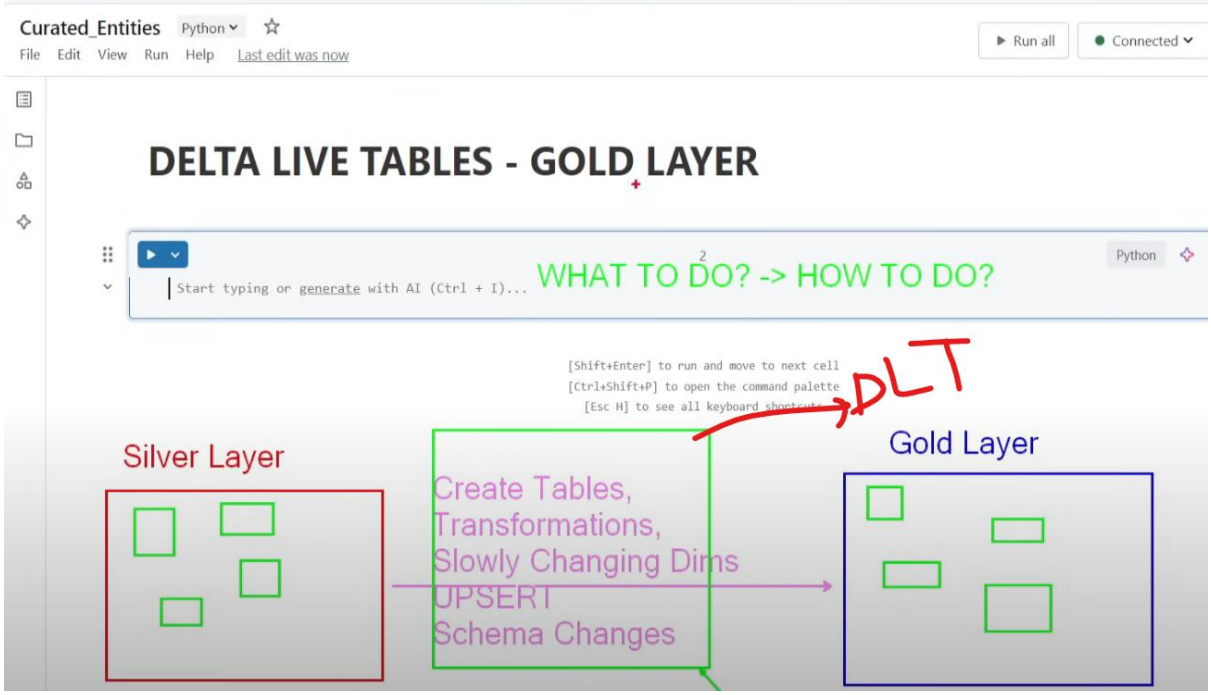


In SQL

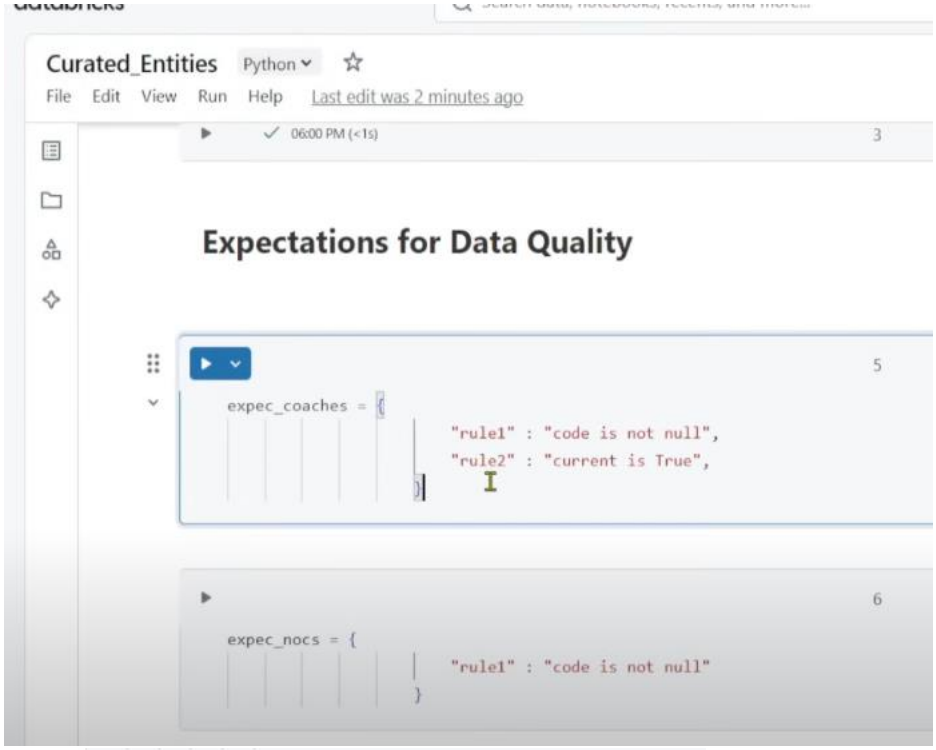


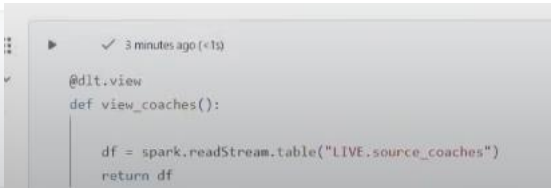
Now we have everything in silver

GOLD



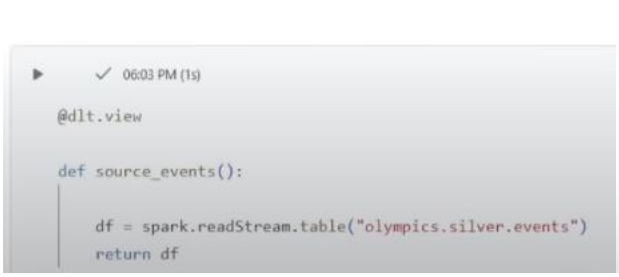
Just tell DLT what to do, it will do How to do i.e. will take care of everything

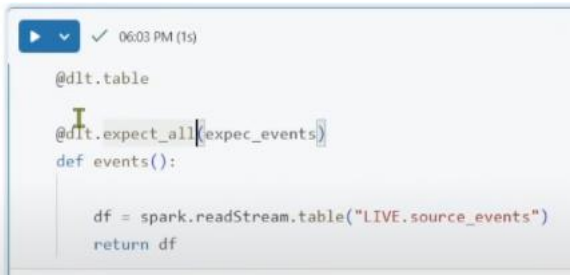




NOCS DLT Pipeline

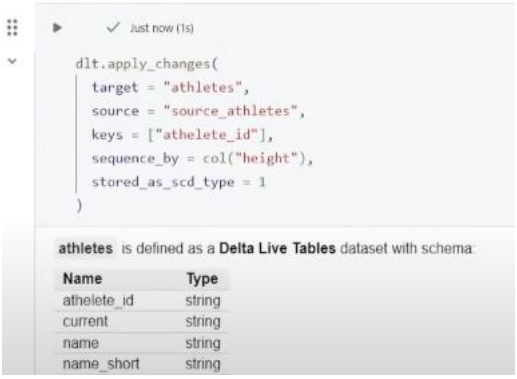
Events DLT Pipeline



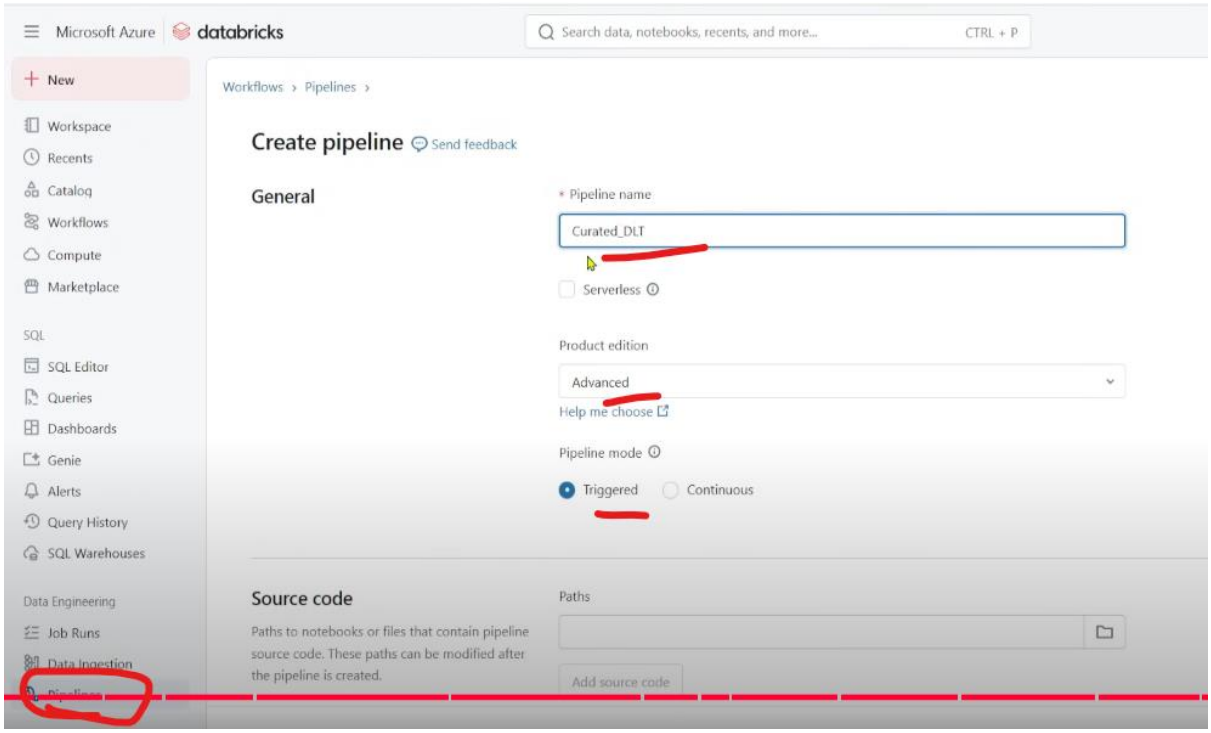


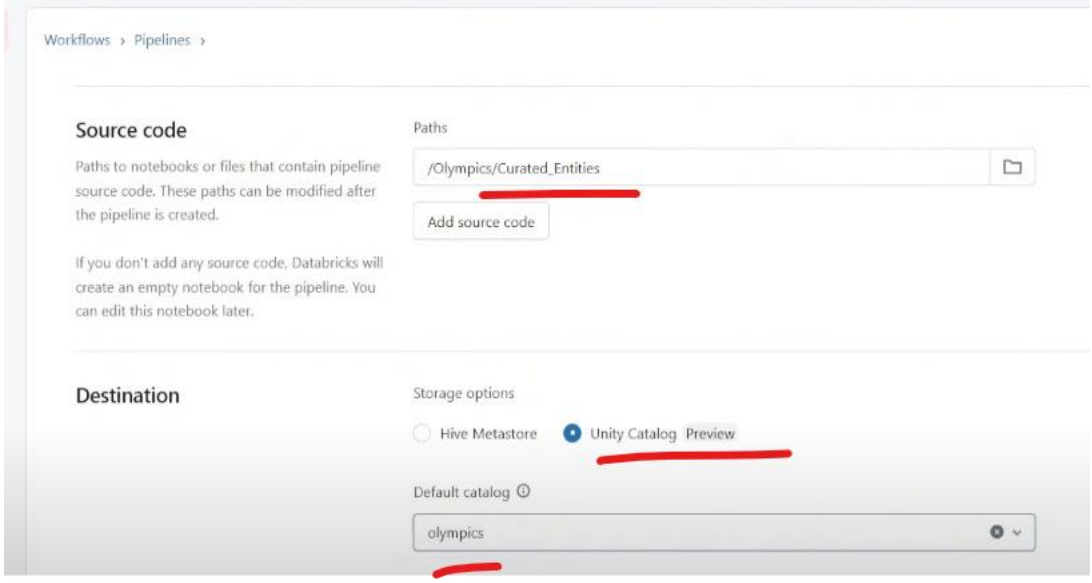
CDC/Apply Changes in DLT

CDC - APPLY CHANGES (DLT)



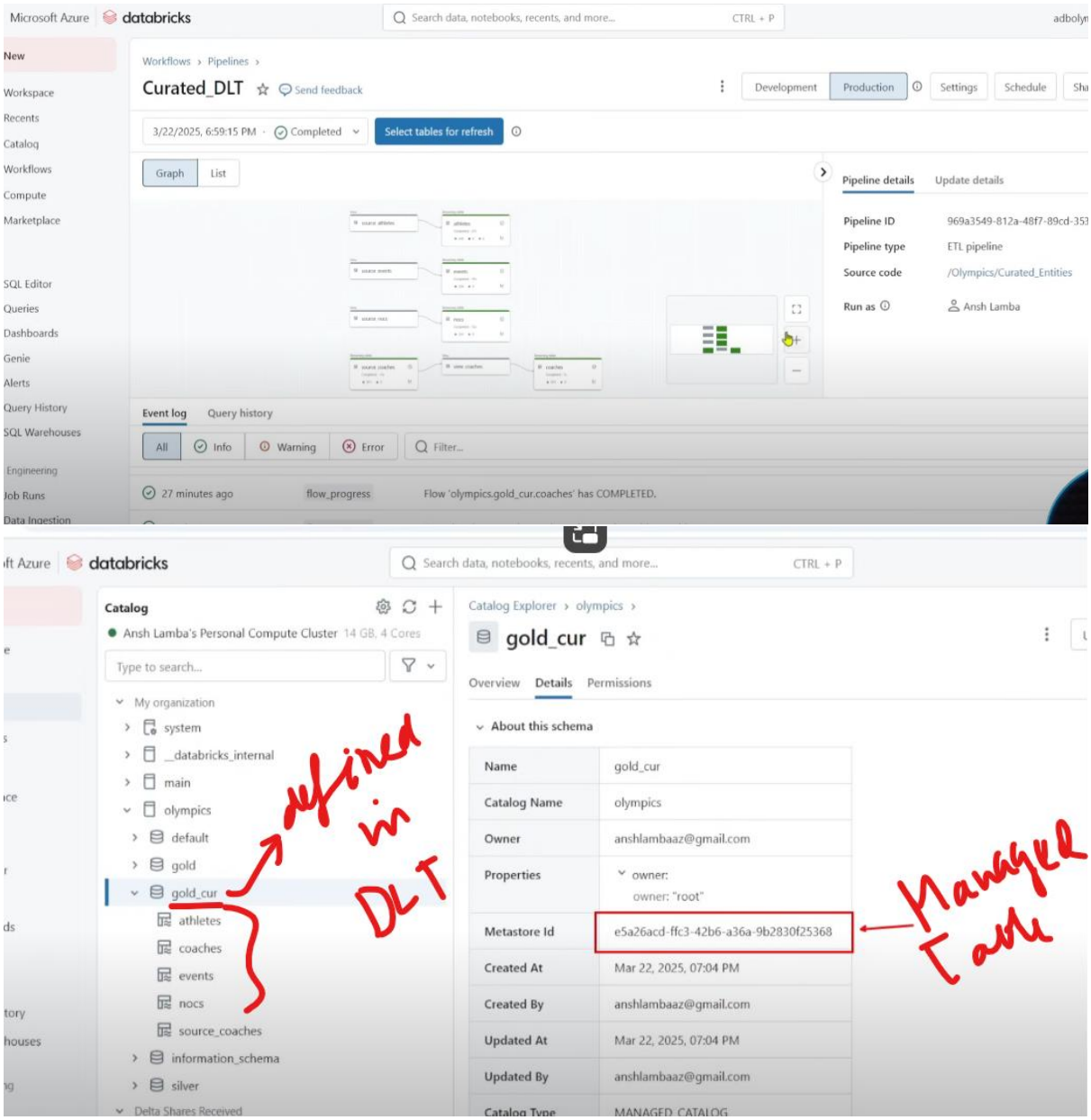
Create a pipeline


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Pipelines
Workflows › Pipelines ›
Create pipeline
Send feedback
General
* Pipeline name
Curated_DLT
Serverless
Product edition
Advanced
Help me choose
Pipeline mode
Triggered
Continuous
Source code
Paths to notebooks or files that contain pipeline source code. These paths can be modified after the pipeline is created.
Paths
Add source code

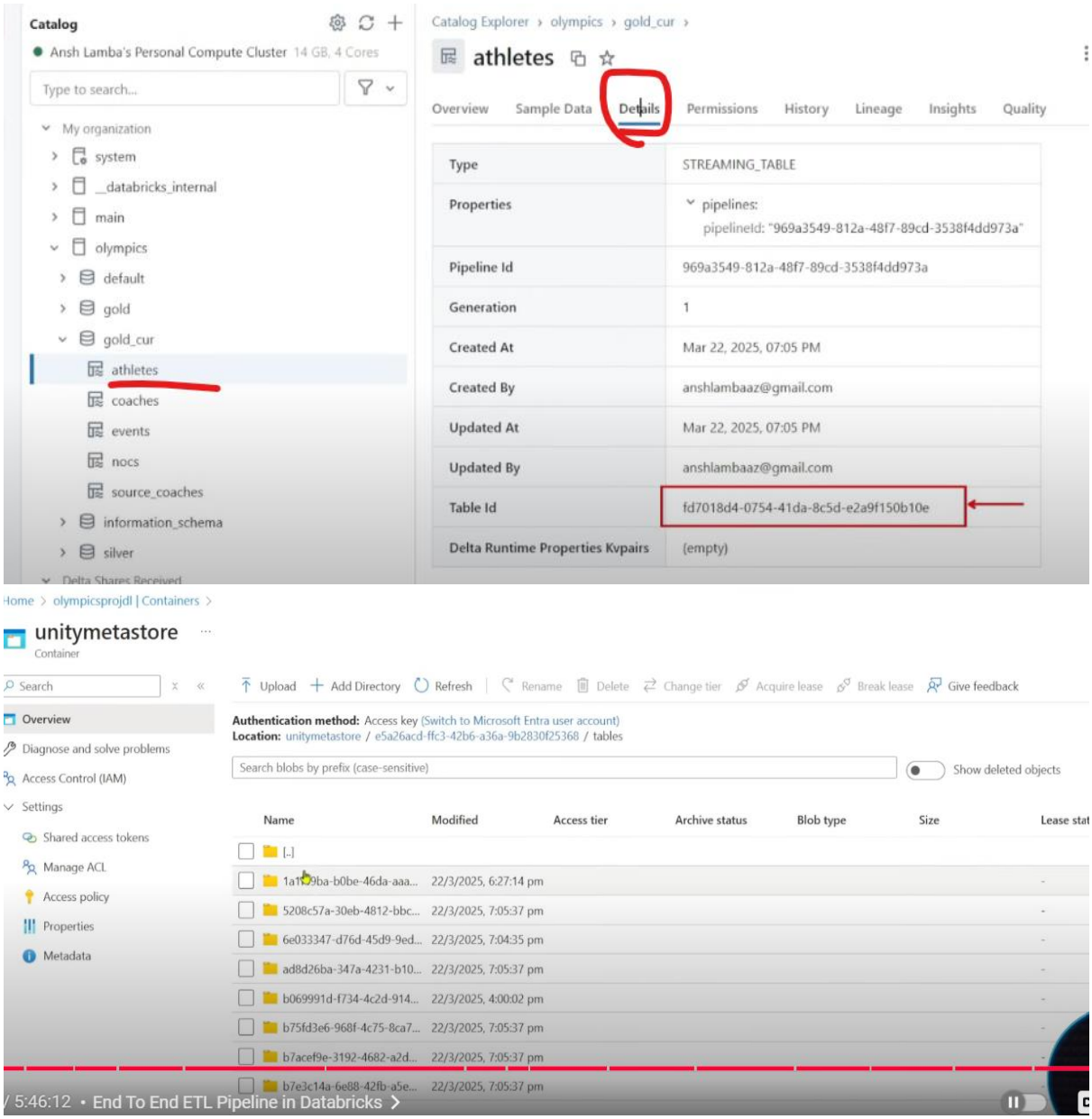

Workflows › Pipelines ›
Source code
Paths to notebooks or files that contain pipeline source code. These paths can be modified after the pipeline is created.
If you don't add any source code, Databricks will create an empty notebook for the pipeline. You can edit this notebook later.
Paths
/Olympics/Curated_Entities
Add source code
Destination
Storage options
Hive Metastore
Unity Catalog Preview
Default catalog
olympics


Default catalog
olympics
* Default schema
gold_cur

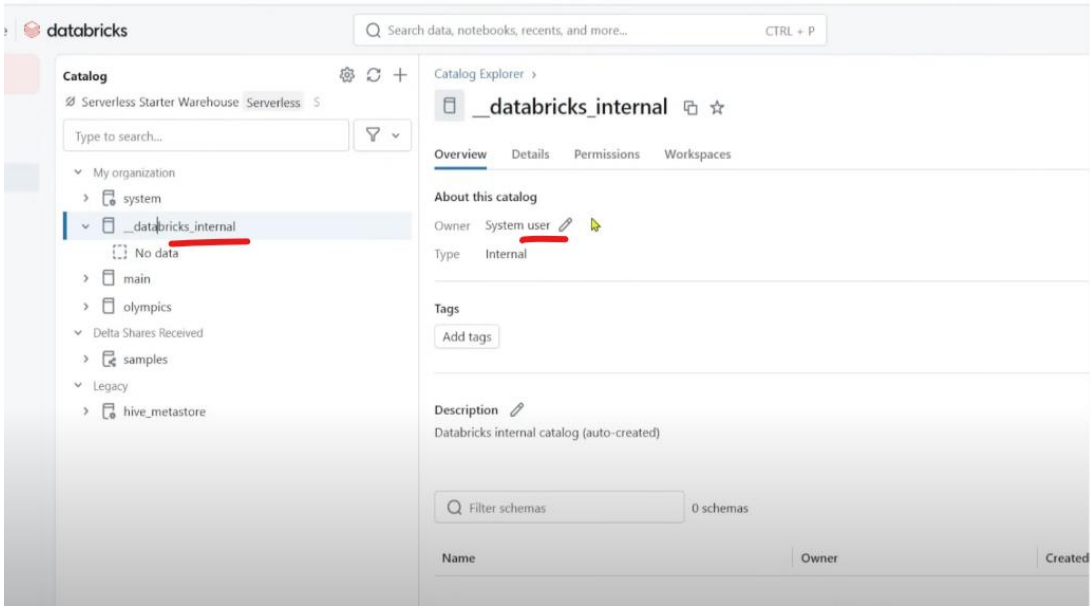
Run the DLT


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
New
Workspace
Recents
Catalog
Workflows
Compute
Marketplace
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses
Engineering
Job Runs
Data Ingestion
Workflows › Pipelines ›
Curated_DLT
Send feedback
Development
Production
Settings
Schedule
3/22/2025, 6:59:15 PM · Completed
Select tables for refresh
Graph
List
Pipeline details
Update details
Pipeline ID
969a3549-812a-48f7-89cd-353
Pipeline type
ETL pipeline
Source code
/Olympics/Curated_Entities
Run as
Ansh Lamba
Event log
Query history
All
Info
Warning
Error
Filter...
27 minutes ago
flow_progress
Flow 'olympics.gold_cur.coaches' has COMPLETED.

databricks
Search data, notebooks, recents, and more...
CTRL + P
Catalog
Ansh Lamba's Personal Compute Cluster 14 GB, 4 Cores
Type to search...
My organization
system
__databricks_internal
main
olympics
default
gold
gold_cur
athletes
coaches
events
nocs
source_coaches
information_schema
silver
Delta Shares Received
defined in DLT
Catalog Explorer › olympics ›
gold_cur
Overview
Details
Permissions
About this schema
Name
gold_cur
Catalog Name
olympics
Owner
anshlambaaz@gmail.com
Properties
owner:
owner: "root"
Metastore Id
e5a26acd-ffc3-42b6-a36a-9b2830f25368
Created At
Mar 22, 2025, 07:04 PM
Created By
anshlambaaz@gmail.com
Updated At
Mar 22, 2025, 07:04 PM
Updated By
anshlambaaz@gmail.com
Catalog Type
MANAGED_CATALOG
Manager Table



These are all Delta Live Tables, No data is found for that table



Everything os managed by databricks_internal which we don't have access to.
So what happens the Table ID in ADB UI is mapped with some random id in ADLS, that's why we are not able to track.