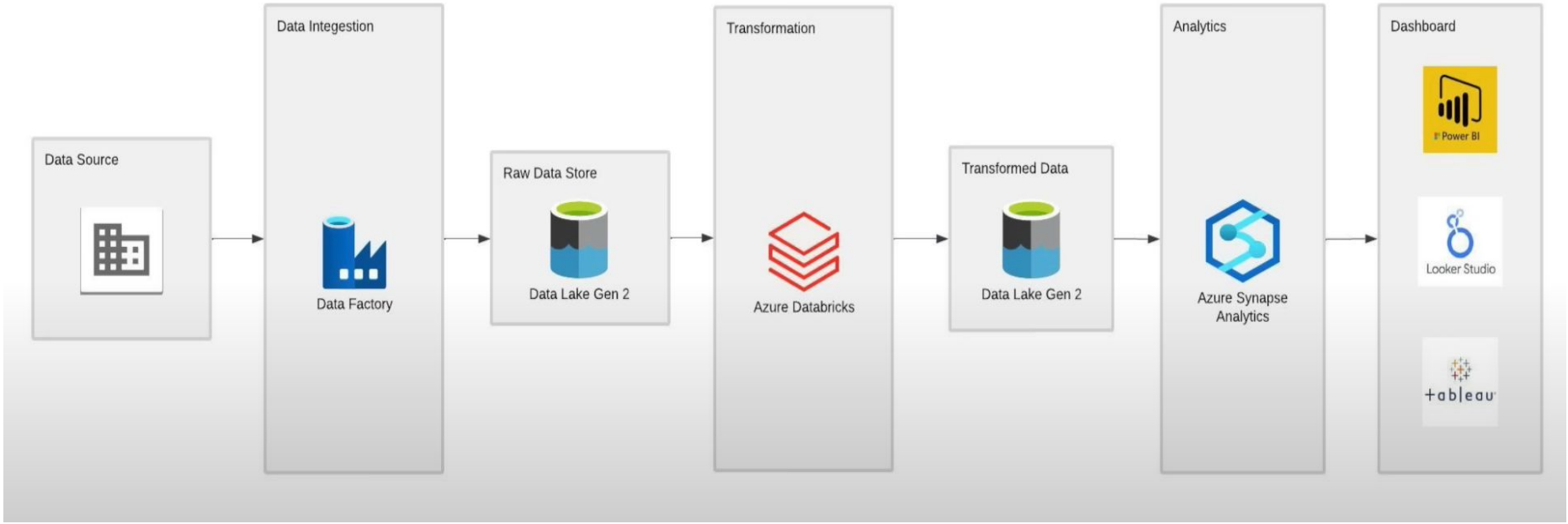
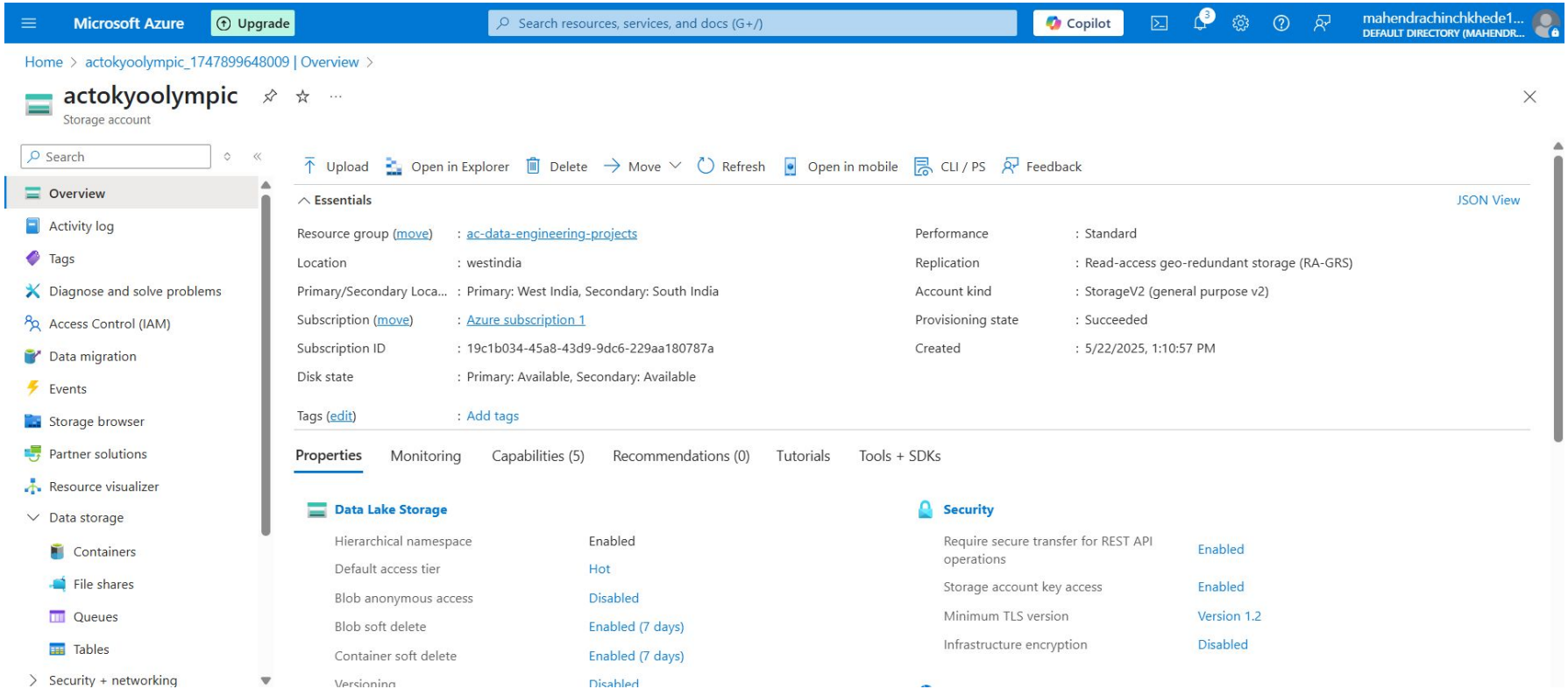
Olympic Data Analytics

Using Microsoft Azure

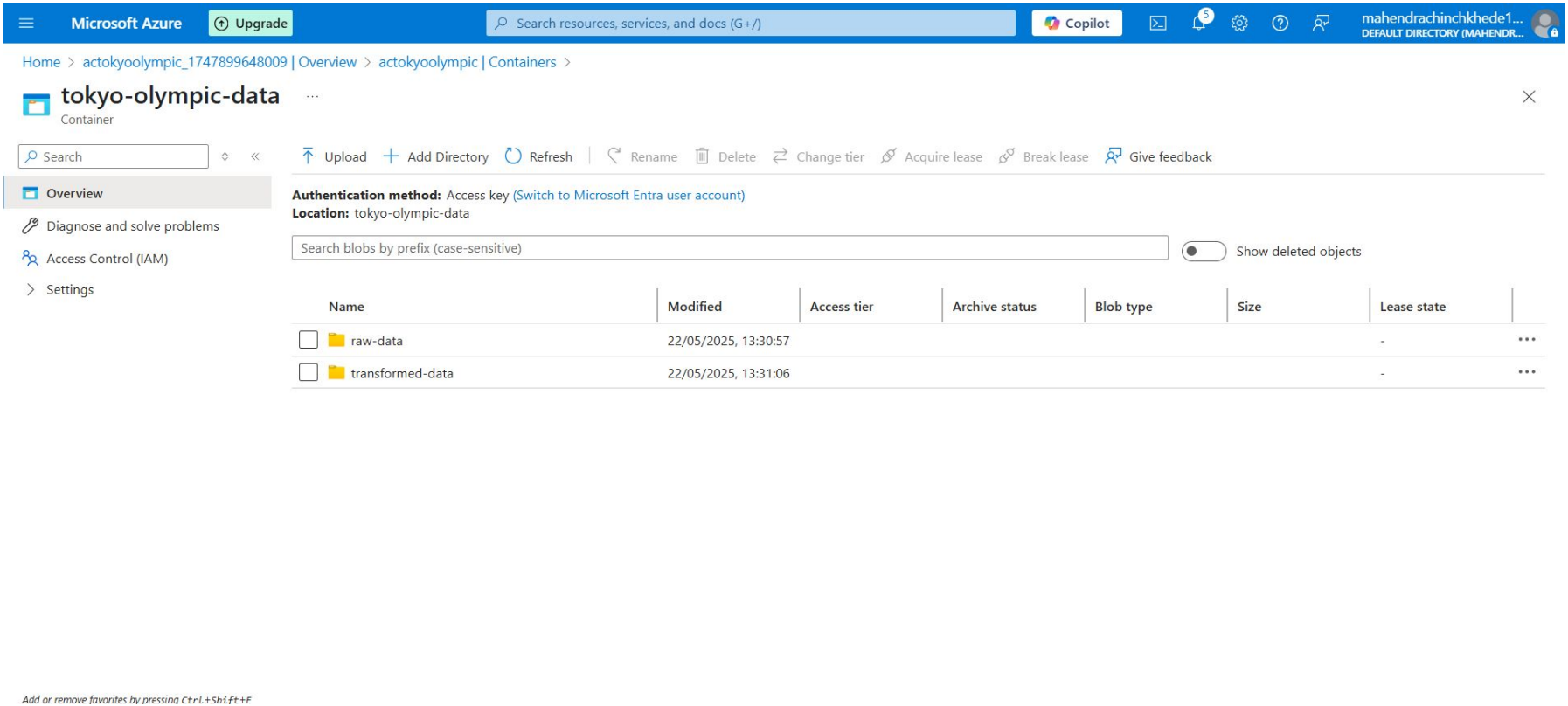
Architecture



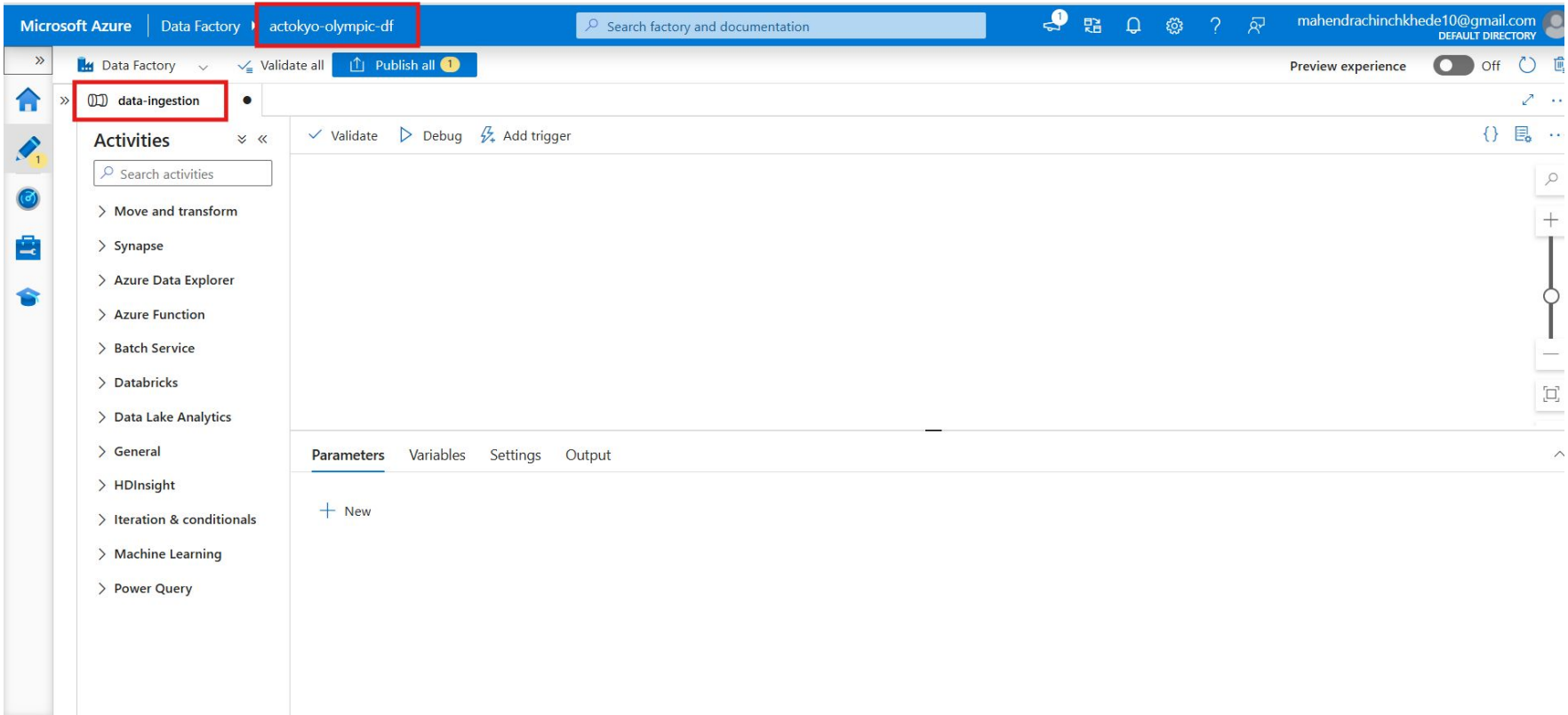
Storage Account



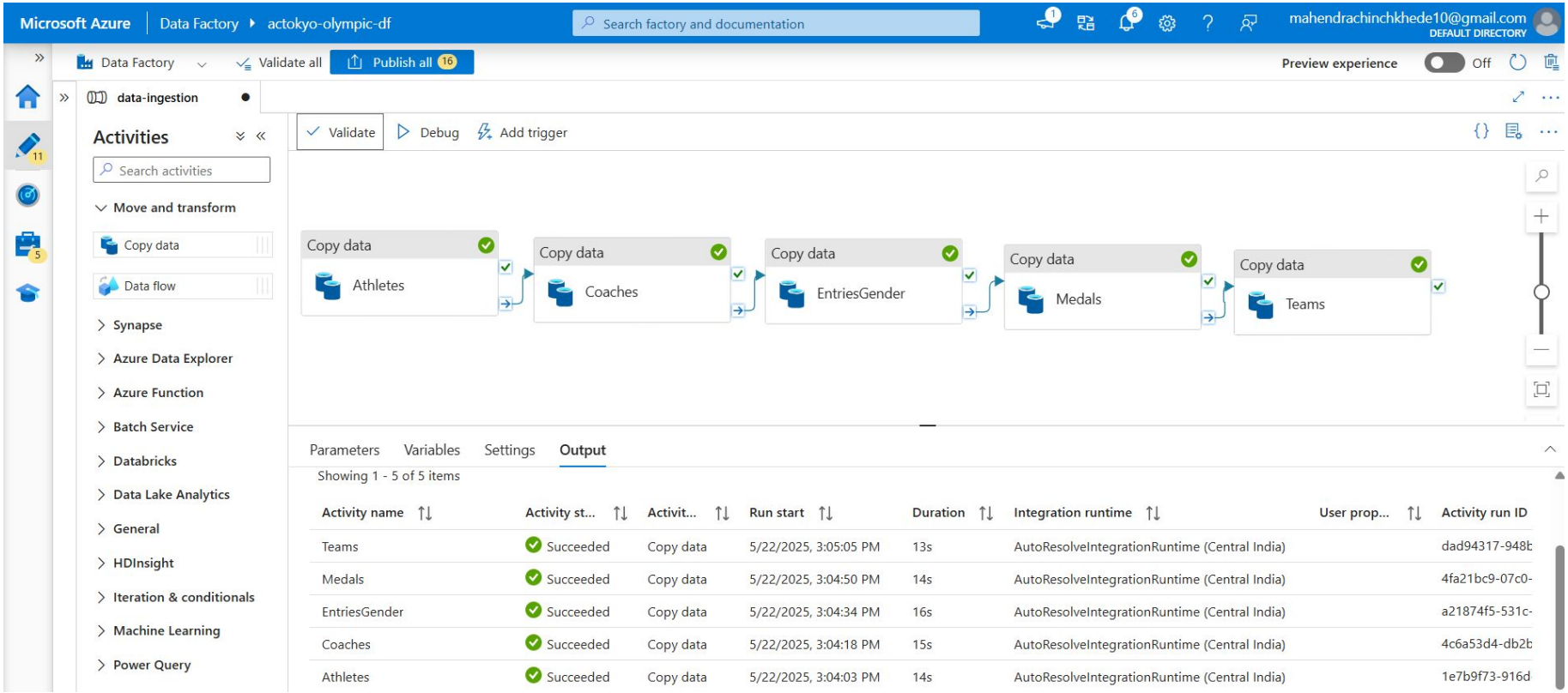
Container

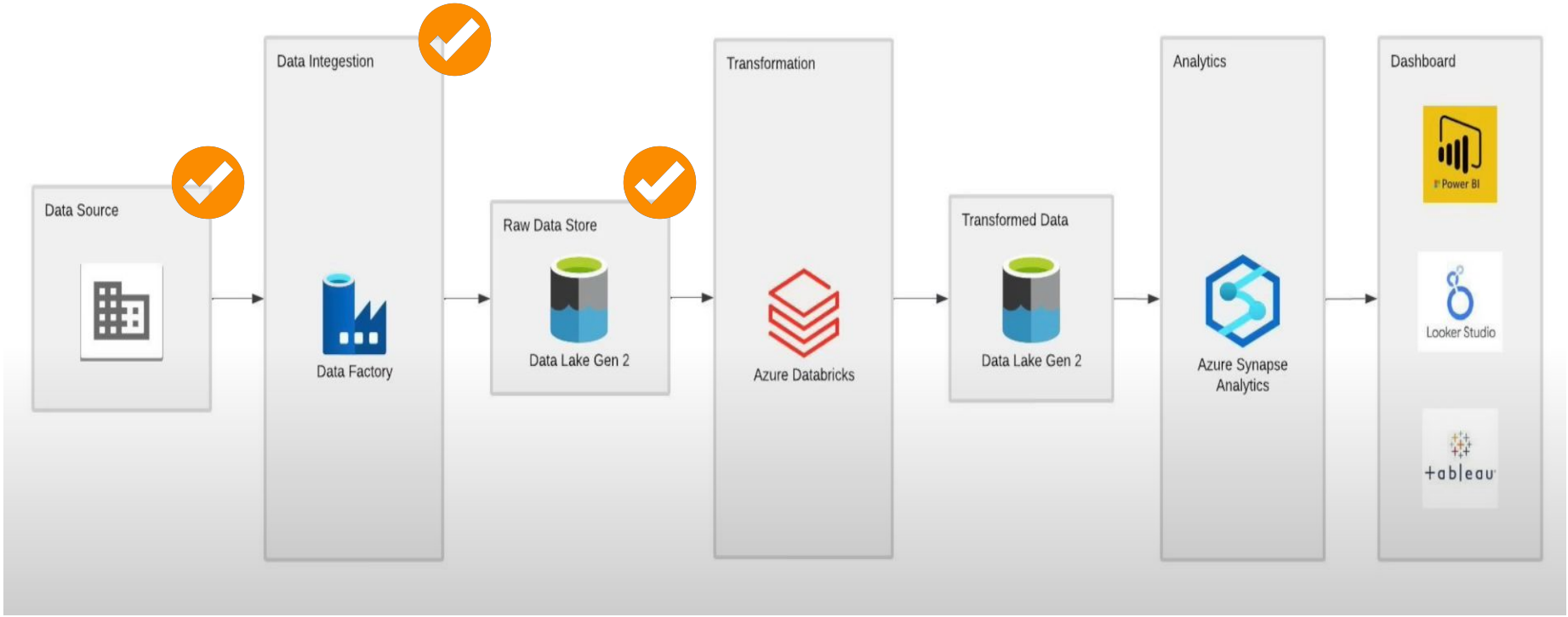

Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Copilot
mahendrachinchkhede1...
DEFAULT DIRECTORY (MAHENDR...
Home > actokyoolympic_1747899648009 | Overview > actokyoolympic | Containers >
tokyo-olympic-data
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: tokyo-olympic-data
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
raw-data
22/05/2025, 13:30:57
transformed-data
22/05/2025, 13:31:06
Add or remove favorites by pressing Ctrl+Shift+F

DataFactory & Pipeline

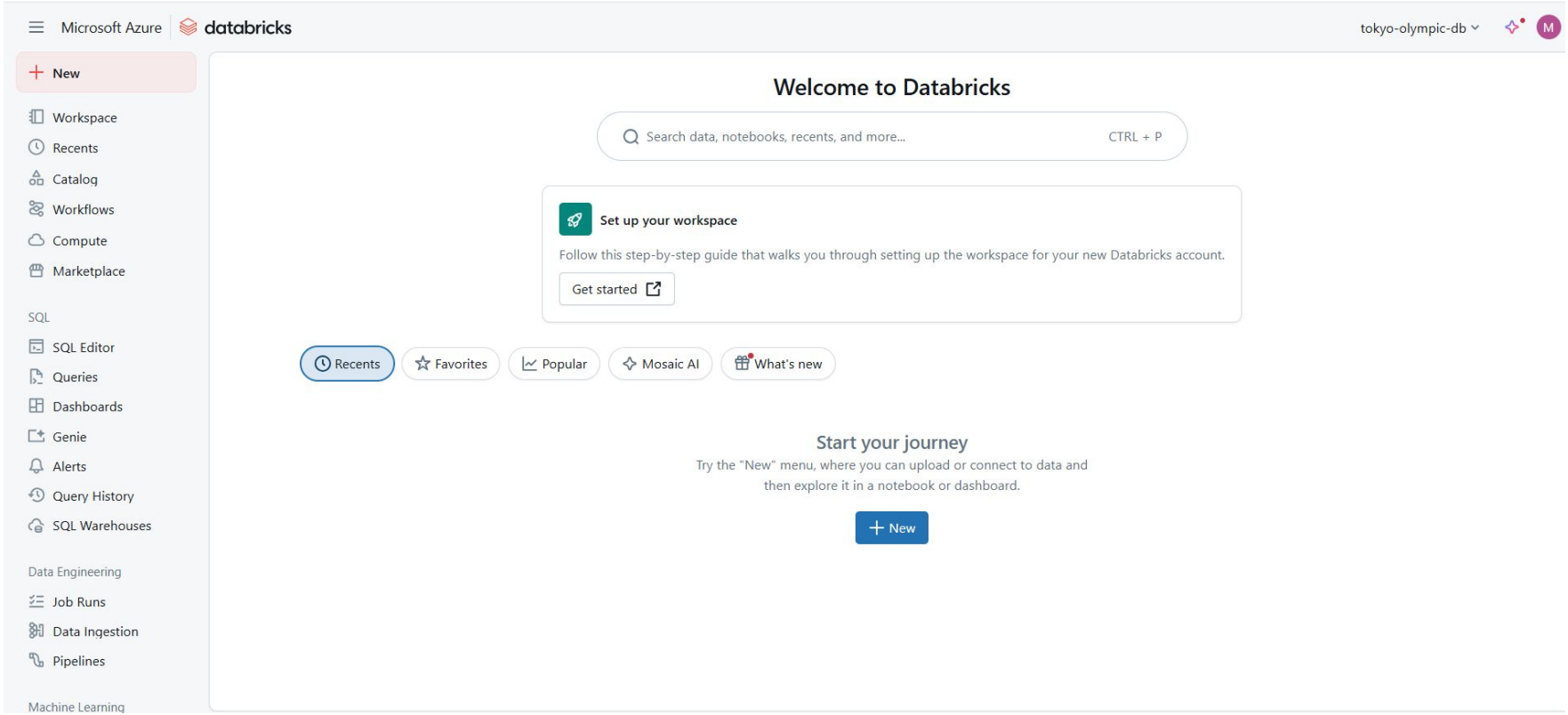


data-ingestion Pipeline

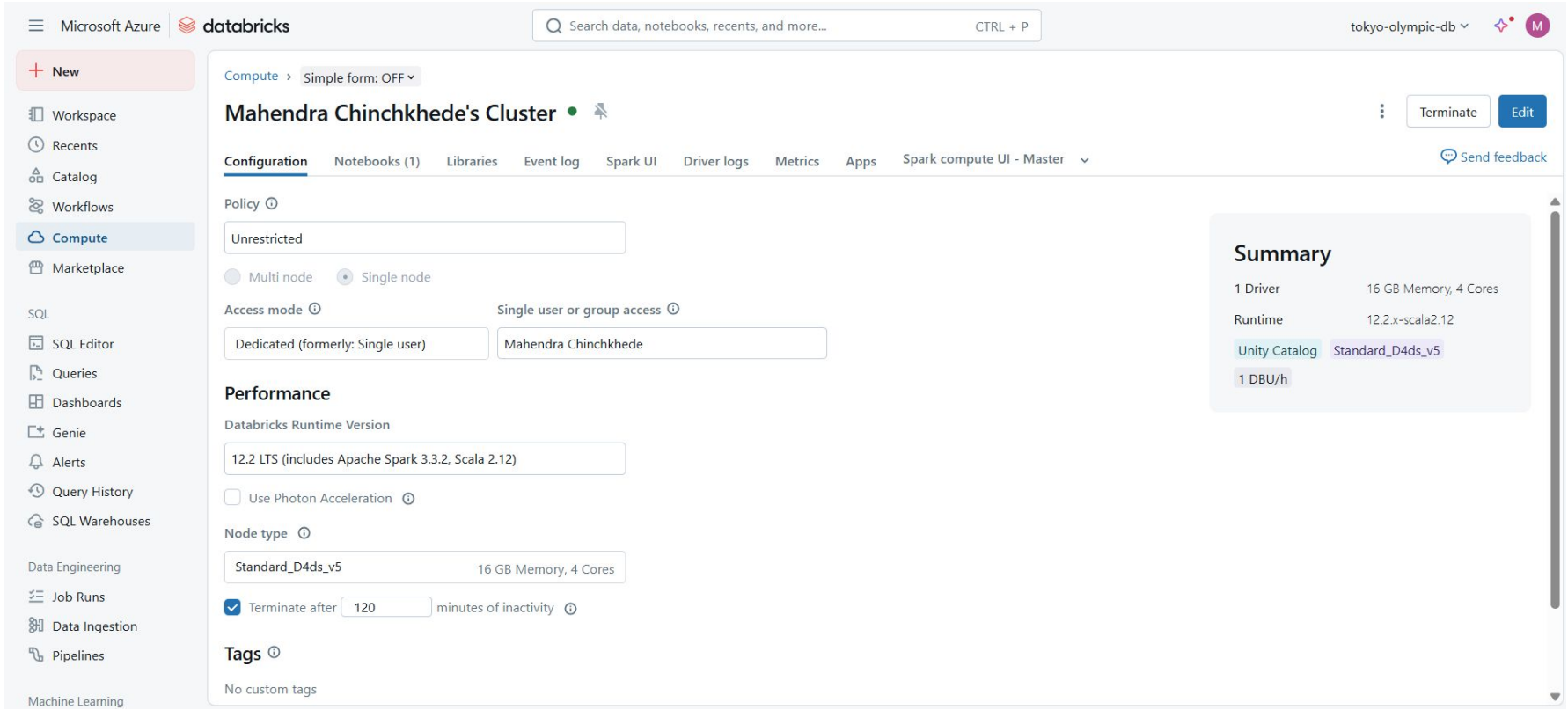



Data Source
Data Integestion
Data Factory
Raw Data Store
Data Lake Gen 2
Transformation
Azure Databricks
Transformed Data
Data Lake Gen 2
Analytics
Azure Synapse Analytics
Dashboard
Power BI
Looker Studio
tableau

Azure Databricks

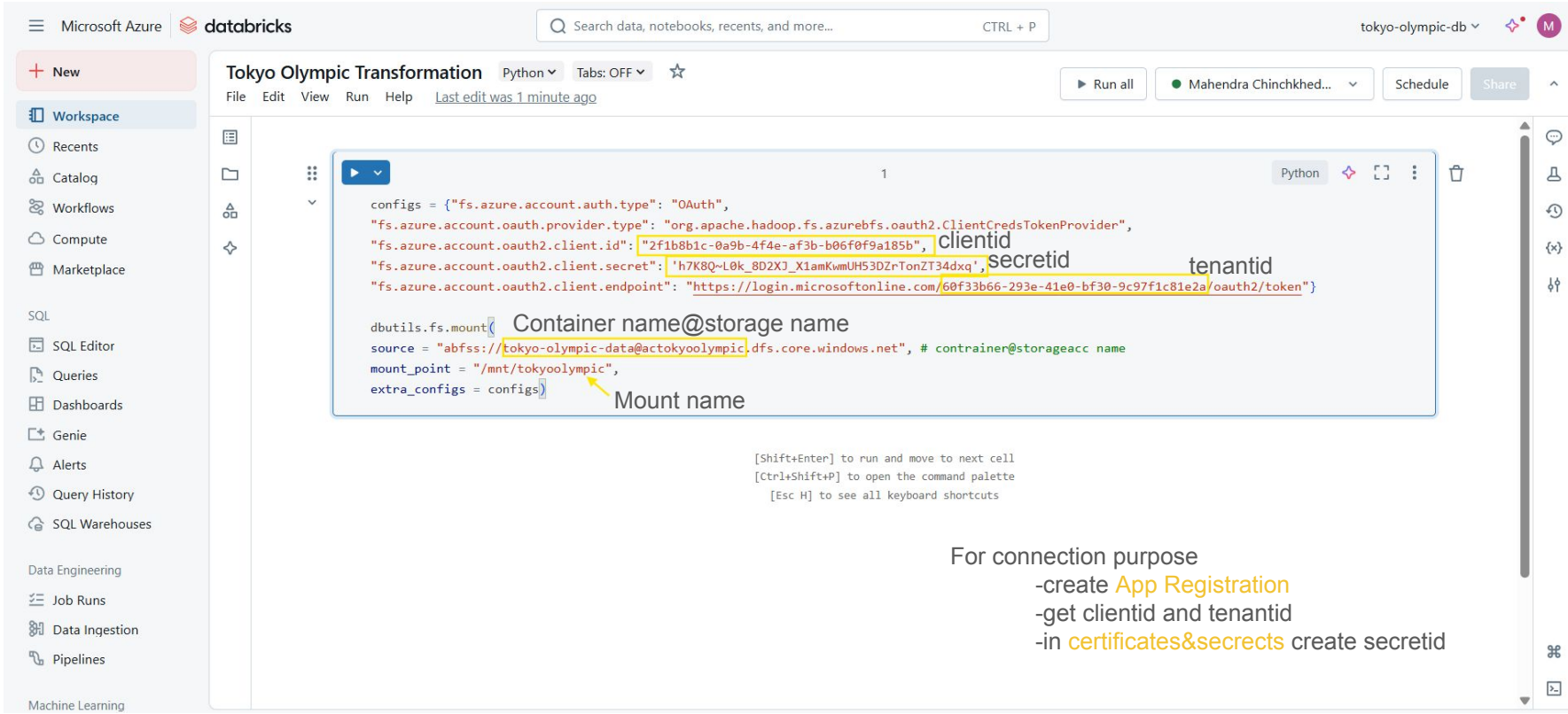


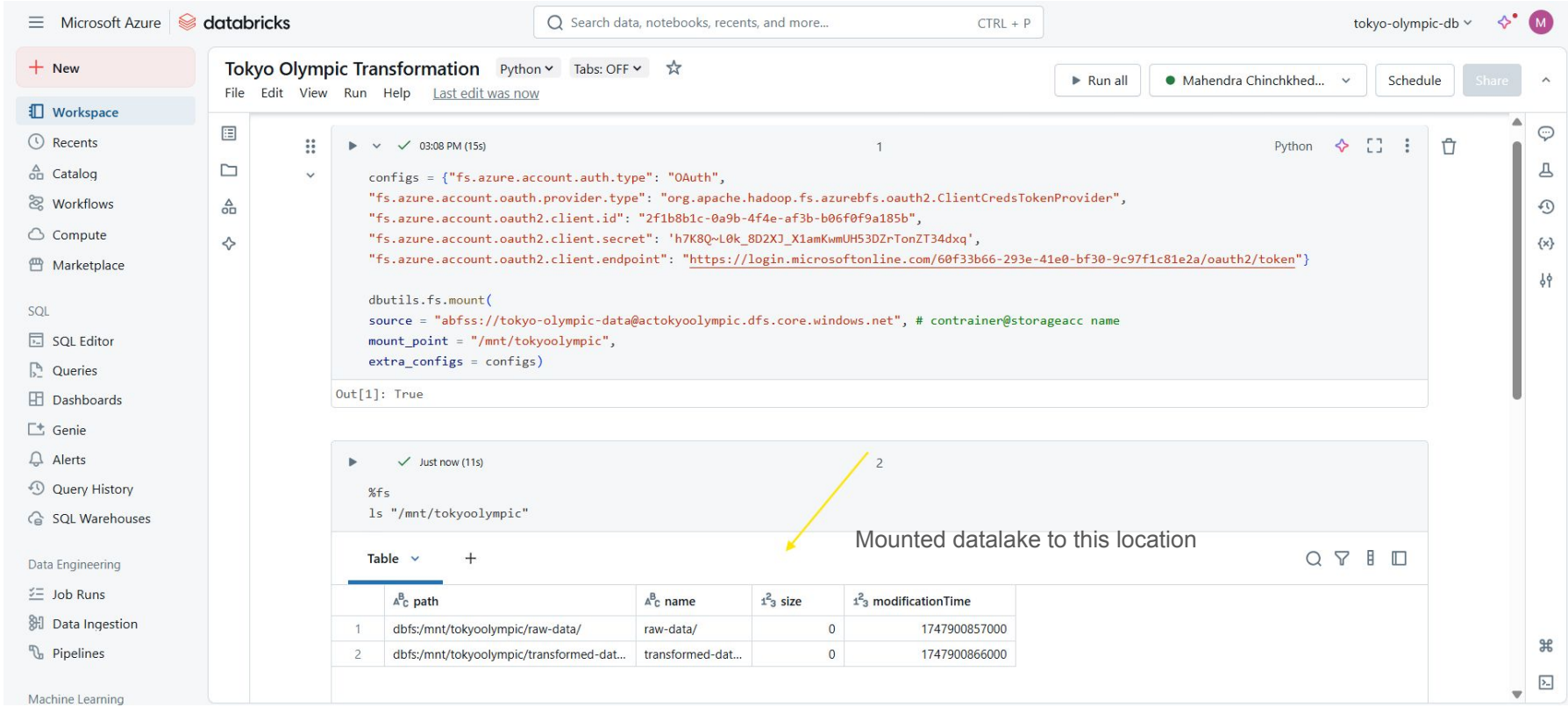
Compute in databricks



Loading data and basic Transformation in Notebook

Notebook - connection & mounting




Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
tokyo-olympic-db
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
Machine Learning
Tokyo Olympic Transformation
Python
Tabs: OFF
File Edit View Run Help
Last edit was now
Run all
Mahendra Chinchkhed...
Schedule
Share
03:08 PM (15s)
1
Python
configs = {"fs.azure.account.auth.type": "OAuth",
"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
"fs.azure.account.oauth2.client.id": "2f1b8b1c-0a9b-4f4e-af3b-b06f0f9a185b",
"fs.azure.account.oauth2.client.secret": 'h7K8Q~L0k_8D2XJ_X1amKwmUH53DZrTonZT34dxq',
"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/60f33b66-293e-41e0-bf30-9c97f1c81e2a/oauth2/token"}
dbutils.fs.mount(
source = "abfss://tokyo-olympic-data@actokyoolympic.dfs.core.windows.net", # contrainer@storageacc name
mount_point = "/mnt/tokyoolympic",
extra_configs = configs)
Out[1]: True
Just now (11s)
2
%fs
ls "/mnt/tokyoolympic"
Mounted datalake to this location
Table
path
name
size
modificationTime
1 dbfs:/mnt/tokyoolympic/raw-data/ raw-data/ 0 1747900857000
2 dbfs:/mnt/tokyoolympic/transformed-dat... transformed-dat... 0 1747900866000

Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

tokyo-olympic-db

New

Workspace
Recents
Catalog
Workflows
Compute
Marketplace

SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History
SQL Warehouses

Data Engineering
Job Runs
Data Ingestion

Tokyo Olympic Transformation

Python | Tabs: OFF

File Edit View Run Help Last edit was 1 minute ago

Run all | Mahendra Chinchkhed... | Schedule | Share

2 rows | 10.74s runtime — Refreshed 10 minutes ago

03:18 PM (<1s) 3 Python

```
spark
```

SparkSession - hive

SparkContext

Spark UI

Version
v3.3.2
Master
local[*, 4]
AppName
Databricks Shell

Spark session active

1 minute ago (<1s) 4

```
athletes = spark.read.format("csv").option("header","true").load("/mnt/tokyoolympic/raw-data/athletes.csv")
```

(1) Spark Jobs

athletes: pyspark.sql.dataframe.DataFrame = [PersonName: string, Country: string ... 1 more field]

Loading athletes.csv file

3 minutes ago (<1s) 5 Python

```
athletes.show()
```

(1) Spark Jobs

Records fetched successfully

PersonName	Country	Discipline
AALERUD Katrine	Norway	Cycling Road
ABAD Nestor	Spain	Artistic Gymnastics
ABAGNALE Giovanni	Italy	Rowing
ABALDE Alberto	Spain	Basketball
ABALDE Tamara	Spain	Basketball
ABALO Luc	France	Handball
ABAROA Cesar	Chile	Rowing
ABASS Abobakr	Sudan	Swimming
ABBASALI Hamideh	Islamic Republic ...	Karate
ABBASOV Islam	Azerbaijan	Wrestling
ABBINGH Lois	Netherlands	Handball
ABBOT Emily	Australia	Rhythmic Gymnastics
ABBOTT Monica	United States of ...	Baseball/Softball
ABDALLA Abubaker ...	Qatar	Athletics
ABDALLA Maryam	Egypt	Artistic Swimming
ABDALLAH Shahd	Egypt	Artistic Swimming
ABDALRASOOL Mohamed	Sudan	Judo
ABDEL LATIF Radwa	Egypt	Shooting

Just now (<1s)

```
entriesgender.show()
```

(1) Spark Jobs

```
+--------------------+------+----+-----+
|          Discipline|Female|Male|Total|
+--------------------+------+----+-----+
|      3x3 Basketball|    32|  32|   64|
|             Archery|    64|  64|  128|
| Artistic Gymnastics|    98|  98|  196|
|   Artistic Swimming|   105|   0|  105|
|           Athletics|   969|1072| 2041|
|           Badminton|    86|  87|  173|
|   Baseball/Softball|    90| 144|  234|
|          Basketball|   144| 144|  288|
|    Beach Volleyball|    48|  48|   96|
|              Boxing|   102| 187|  289|
|        Canoe Slalom|    41|  41|   82|
|        Canoe Sprint|   123| 126|  249|
|Cycling BMX Frees...|    10|   9|   19|
|  Cycling BMX Racing|    24|  24|   48|
|Cycling Mountain ...|    38|  38|   76|
|        Cycling Road|    70| 131|  201|
|       Cycling Track|    90|  99|  189|
|              Diving|    72|  71|  143|
```

1 minute ago (<1s)

```
entriesgender.printSchema()
```

```
root
 |-- Discipline: string (nullable = true)
 |-- Female: string (nullable = true)
 |-- Male: string (nullable = true)
 |-- Total: string (nullable = true)
```

For entriesgender table, column Female, Male & Total does not have correct datatype

Tokyo Olympic Transformation Python Tabs: OFF

File Edit View Run Help Last edit was now

Run all | Mahendra Chinchkhed... | Schedule

03:45 PM (<1s) 11

```
#We dont have Colums in proper datatypes, so we need to change.
from pyspark.sql.functions import col
from pyspark.sql.types import IntegerType, DateType, DoubleType, BooleanType
```

Just now (<1s) 12

```
entriesgender = entriesgender.withColumn("Female",col("Female").cast(IntegerType()))\
    .withColumn("Male",col("Male").cast(IntegerType()))\
    .withColumn("Total",col("Total").cast(IntegerType()))
```

entriesgender: pyspark.sql.dataframe.DataFrame = [Discipline: string, Female: integer ... 2 more fields]

Converted Female,Male,Total Column datatype into integer

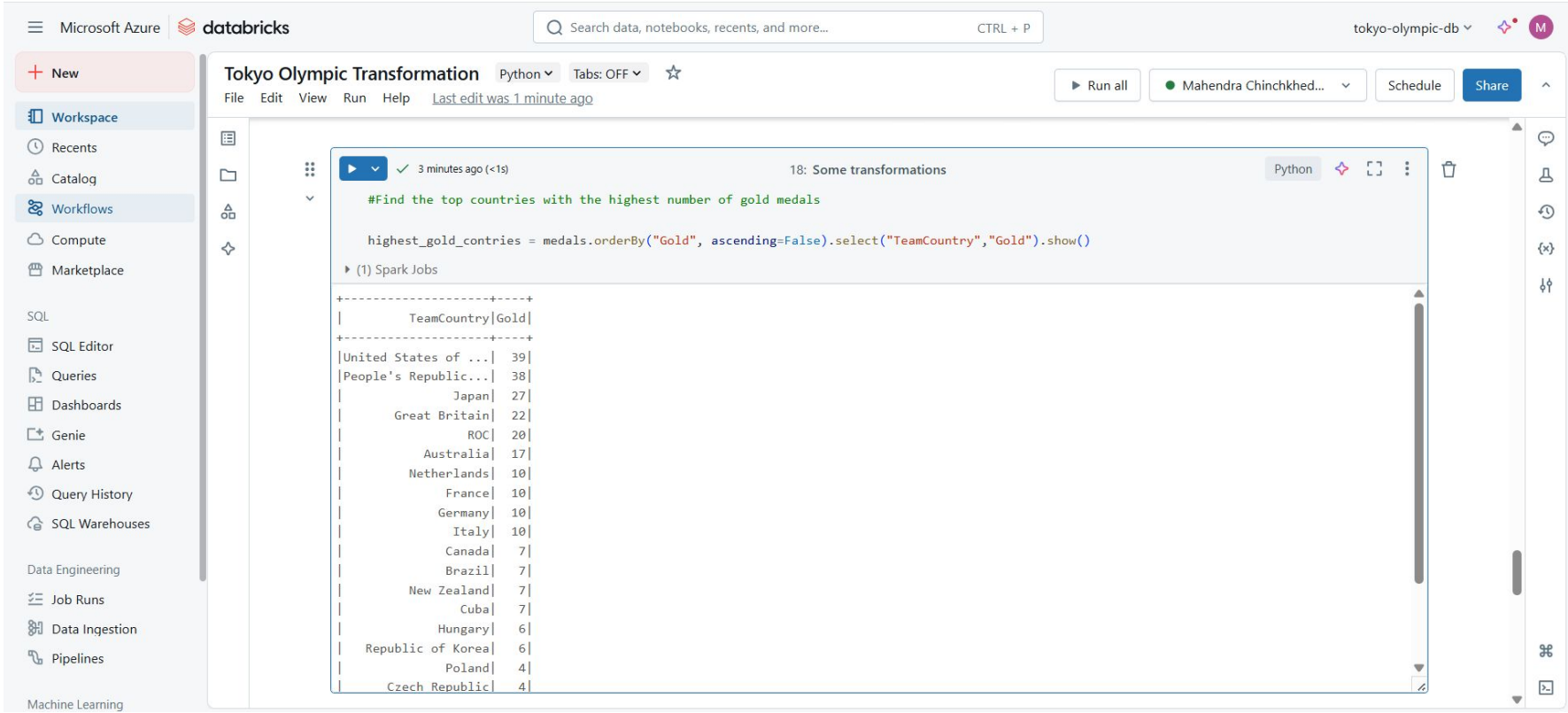
Just now (<1s) 13 Python

```
entriesgender.printSchema()
```

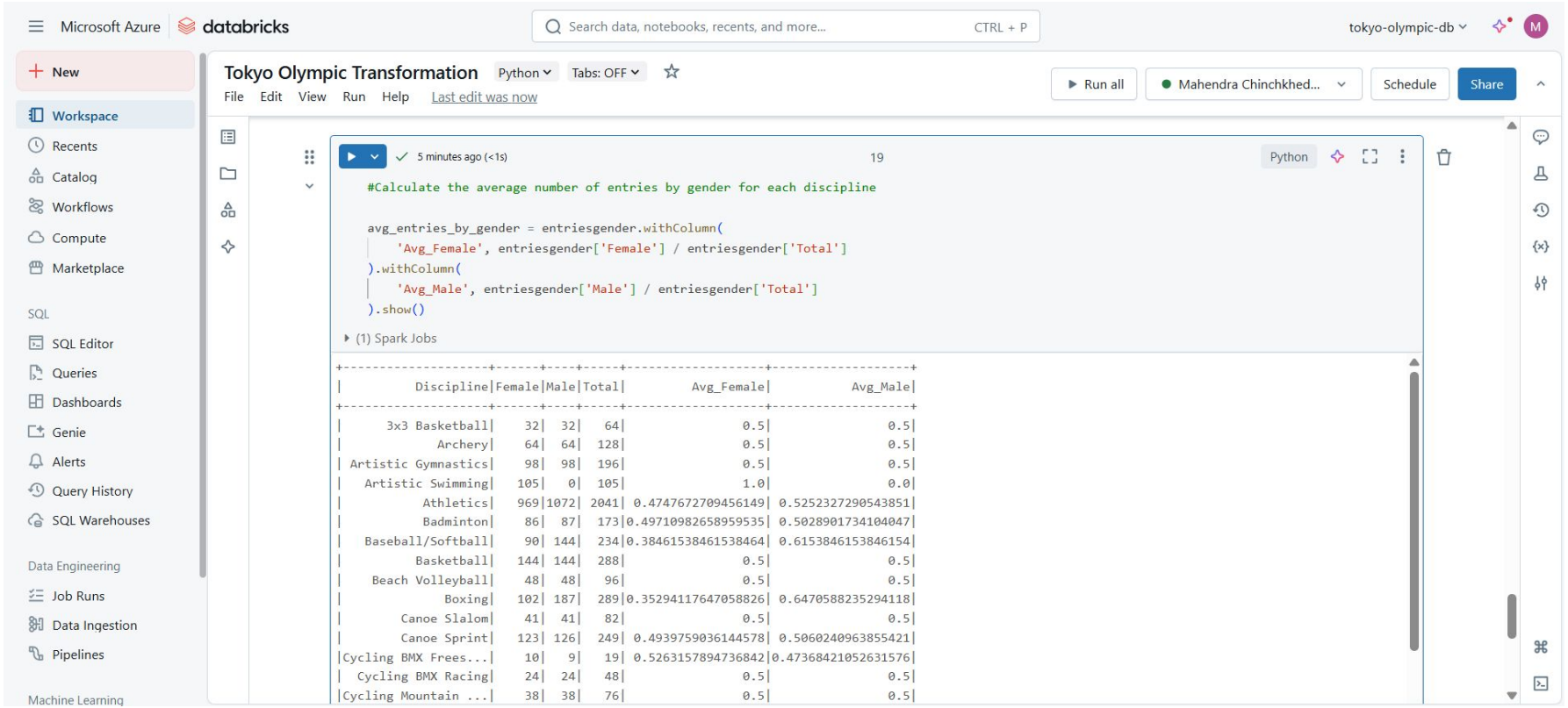
```
root
 |-- Discipline: string (nullable = true)
 |-- Female: integer (nullable = true)
 |-- Male: integer (nullable = true)
 |-- Total: integer (nullable = true)
```

Instead of manually formatting schema, we can use inferSchema function in spark - it will automatically identify datatype and convert accordingly

Find top countries with the highest number of gold medals



Calculate average number of entries by gender for each discipline



Microsoft Azure | databricks

Search data, notebooks, recents, and more... CTRL + P

tokyo-olympic-db

New

Workspace
Recents
Catalog
Workflows
Compute
Marketplace

SQL
SQL Editor
Queries
Dashboards
Genie
Alerts
Query History

Tokyo Olympic Transformation

Python | Tabs: OFF

File Edit View Run Help Last edit was 1 minute ago

Run all | Mahendra Chinchkhed... | Schedule | Share

07:16 PM (<1s) 20

```
#Writing Transformed Data
```

For no file parts

To save changes made in same file

File name

Mount location

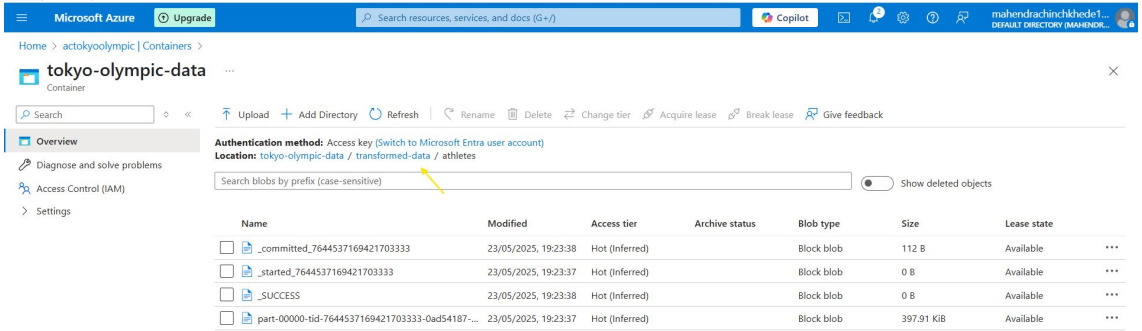
Just now (1s) 21

```
athletes.repartition(1).write.mode("overwrite").option("header",True).csv("/mnt/tokyoolympic/transformed-data/athletes")
```

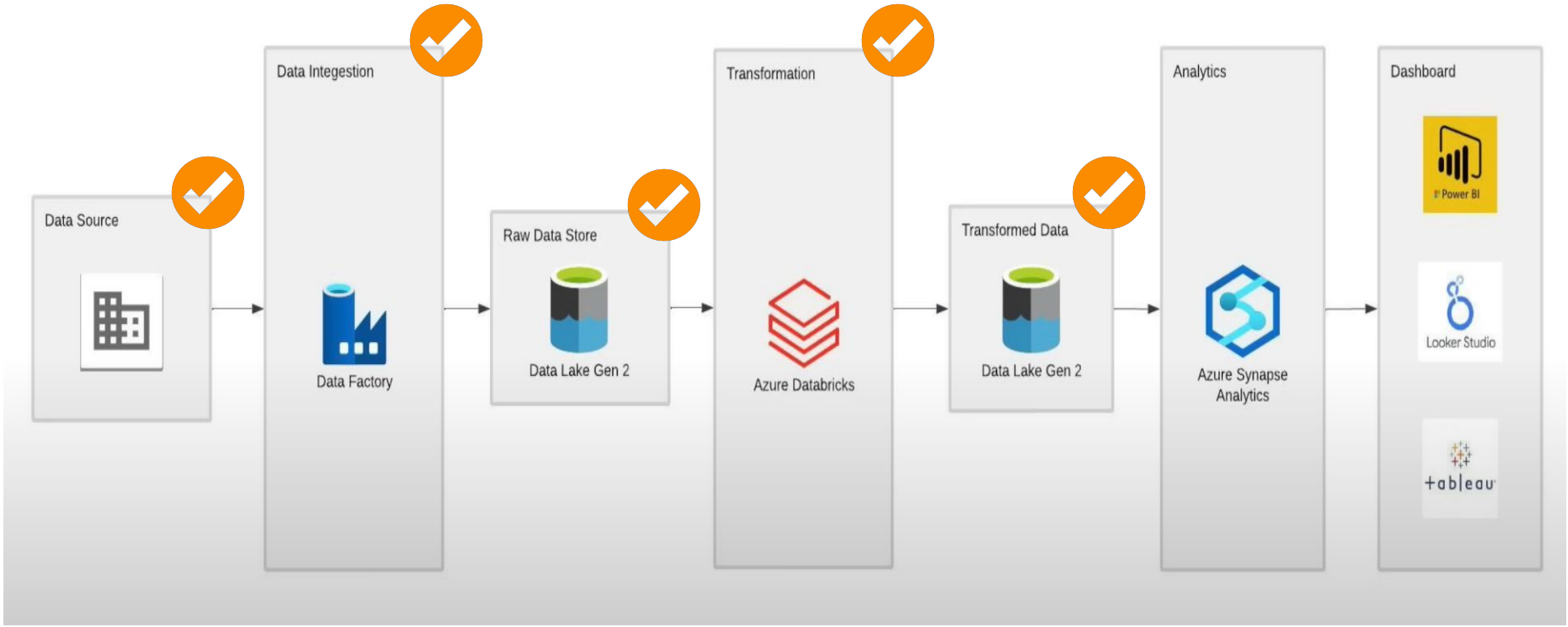
(2) Spark Jobs

22 Python

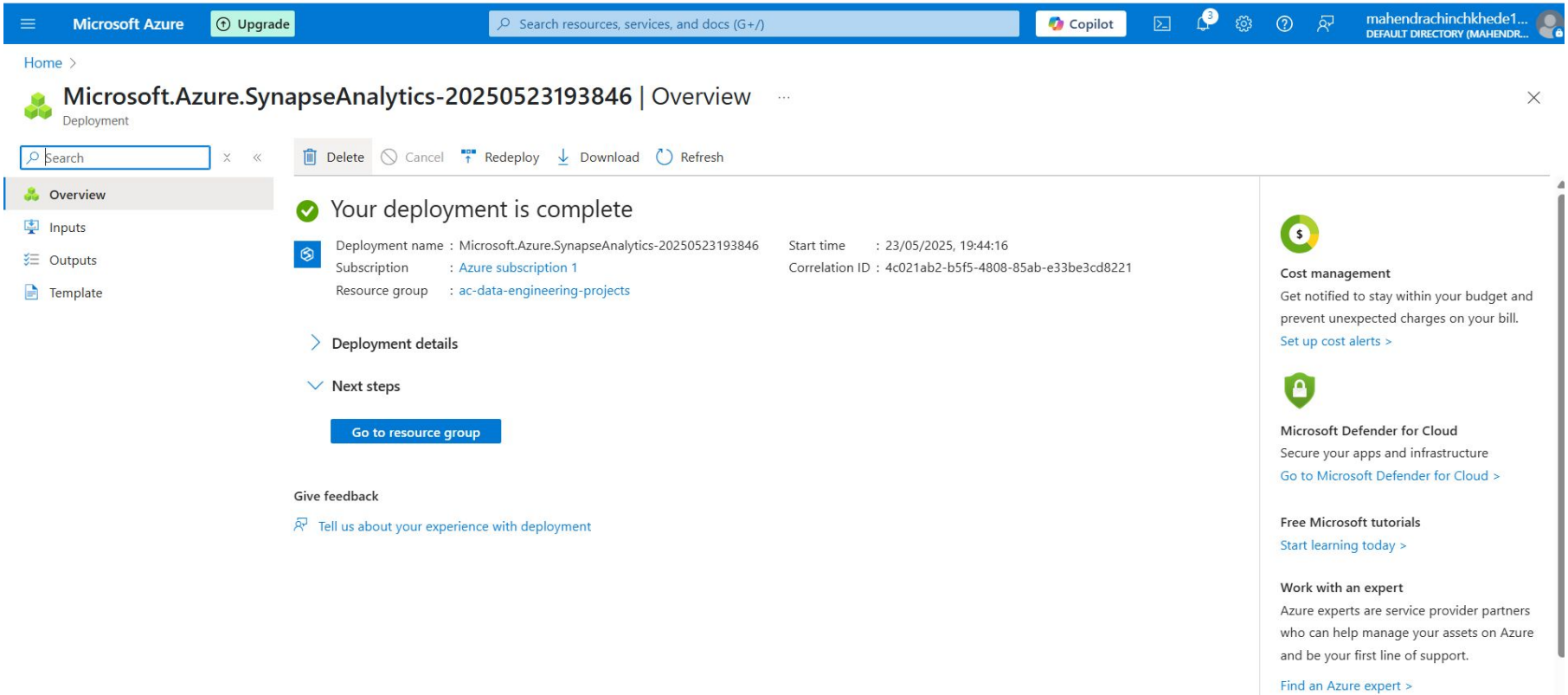
Start typing or generate with AI (Ctrl + I)...



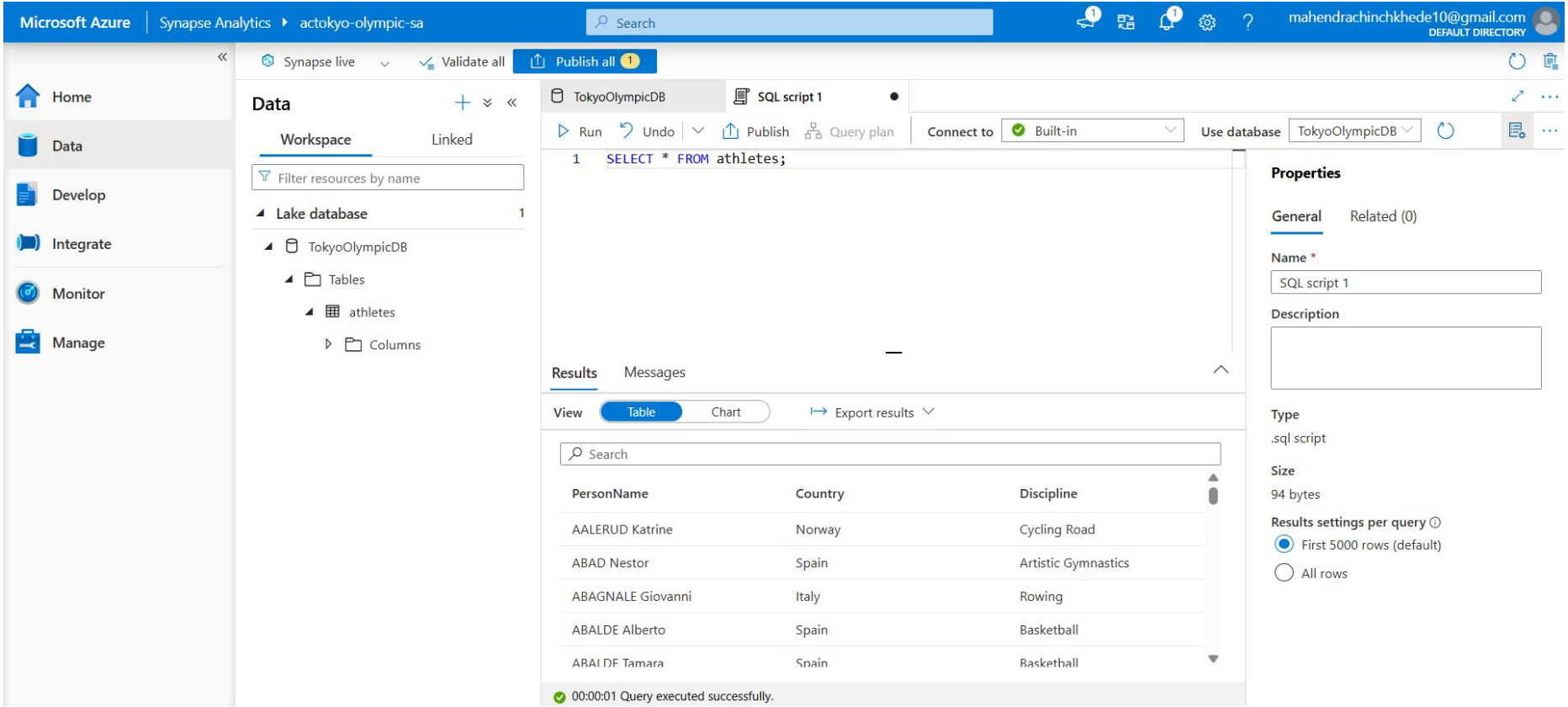
Transformed data saved to desired location successfully


Data Source
Data Integestion
Data Factory
Raw Data Store
Data Lake Gen 2
Transformation
Azure Databricks
Transformed Data
Data Lake Gen 2
Analytics
Azure Synapse Analytics
Dashboard
Power BI
Looker Studio
tableau

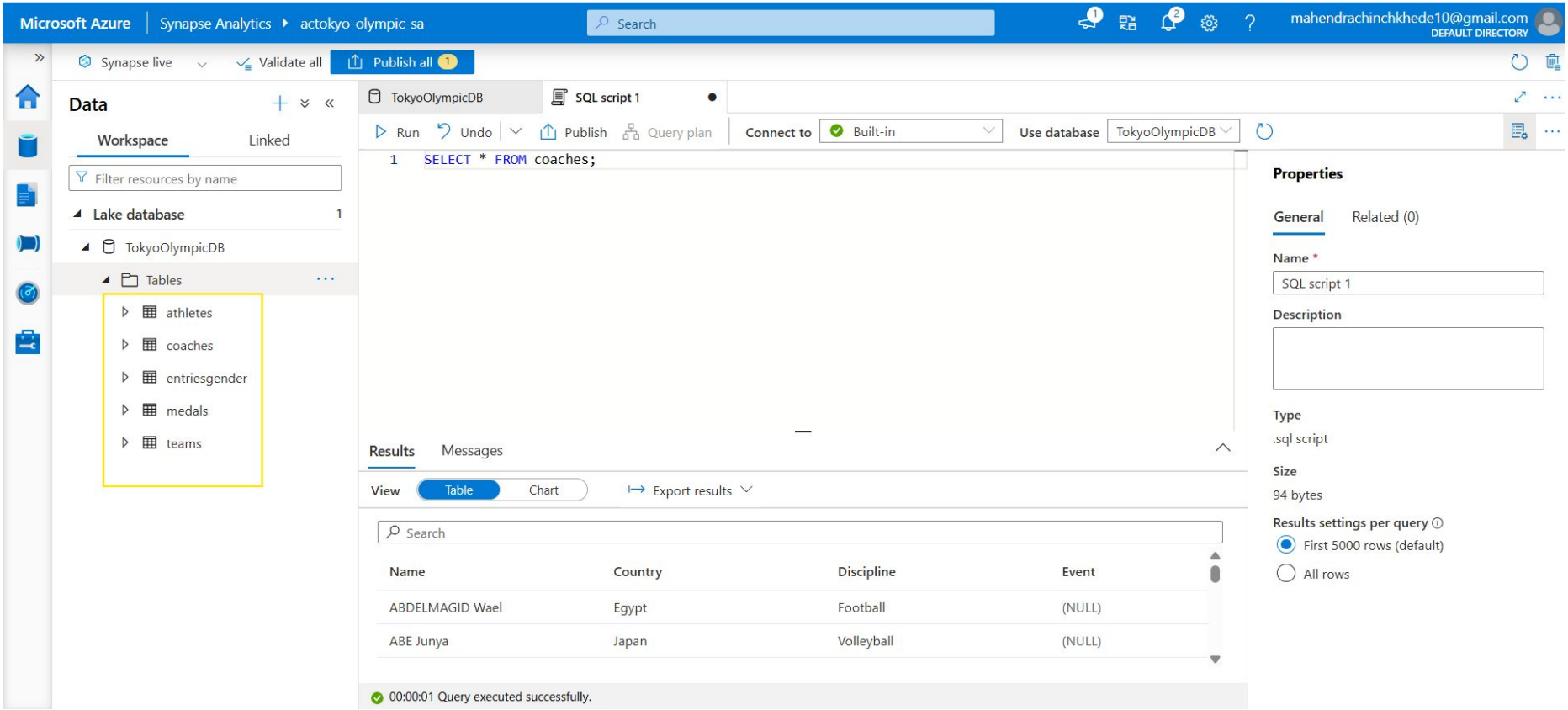
Azure Synapse Analytics



Azure Synapse Analytics

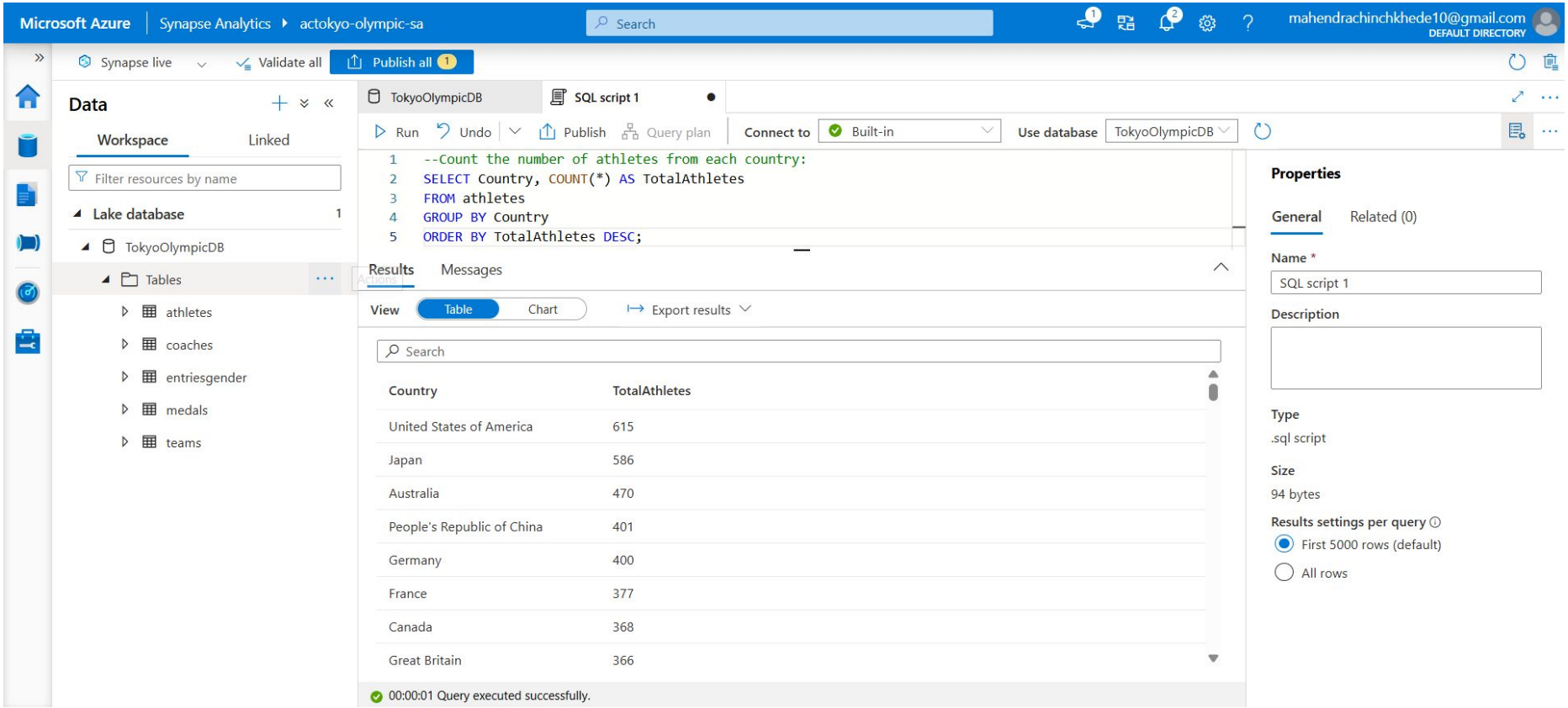


-Created Lake Database (TokyoOlympicDB)
-Imported table athletes from transformed data folder
-performed SQL queries on data



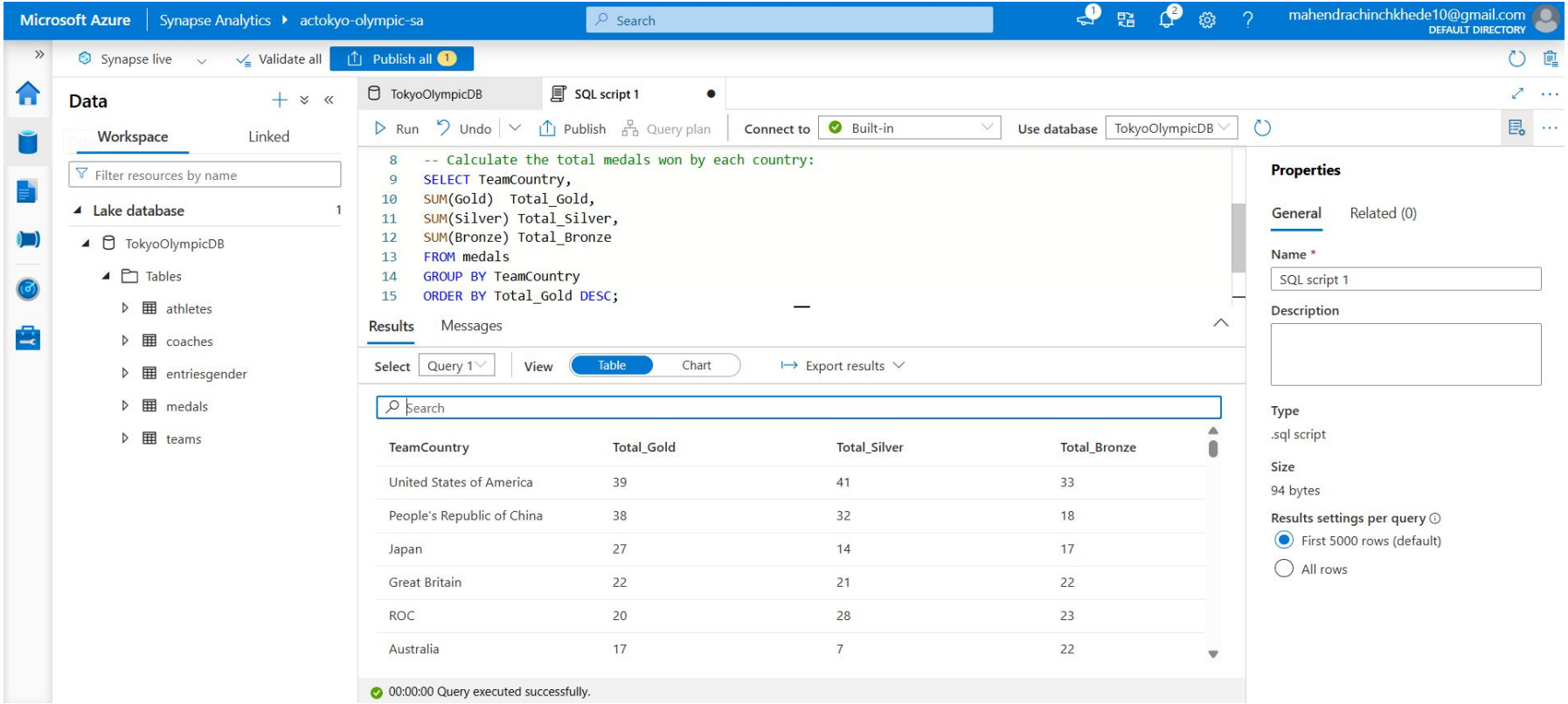
Similarly imported other tables

Count the number of athletes from each country

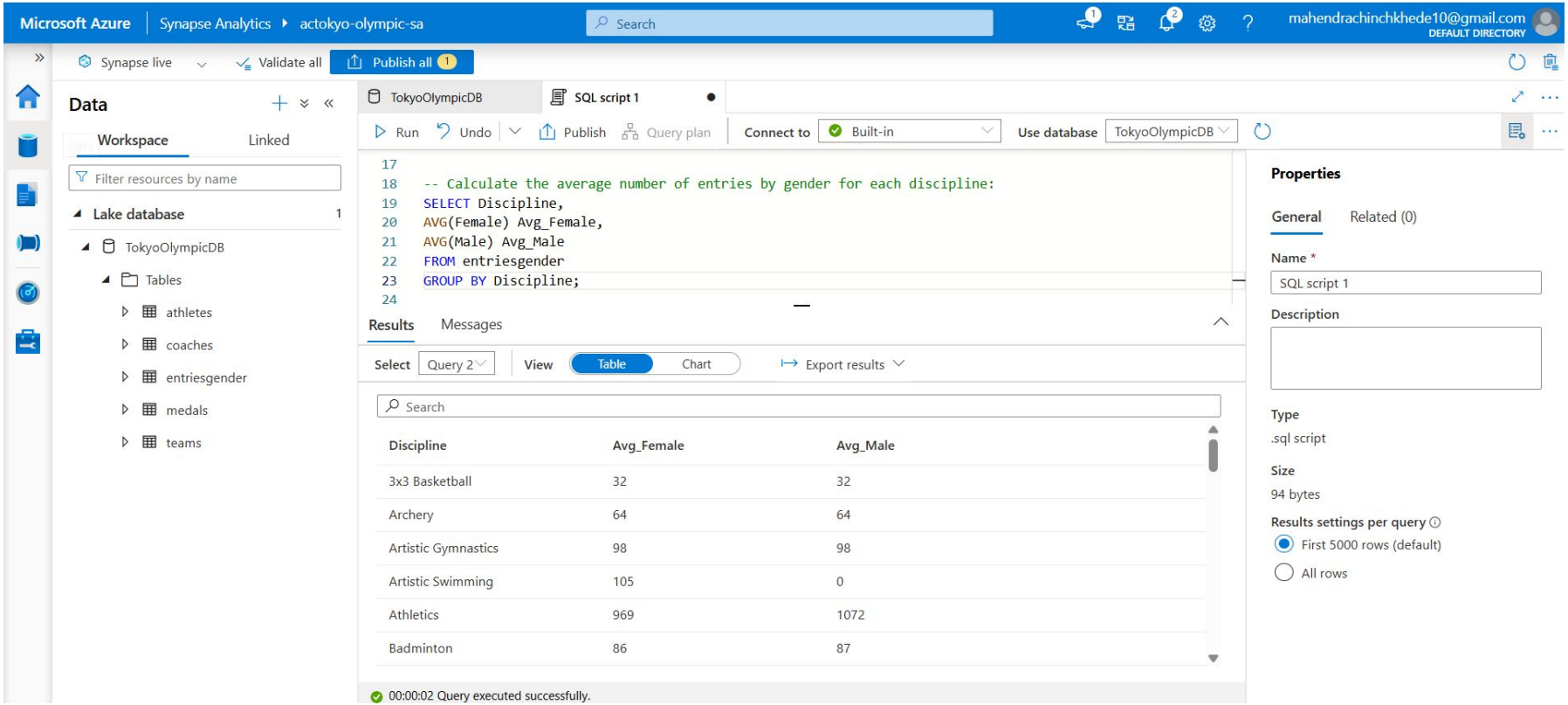


SQL query Operation

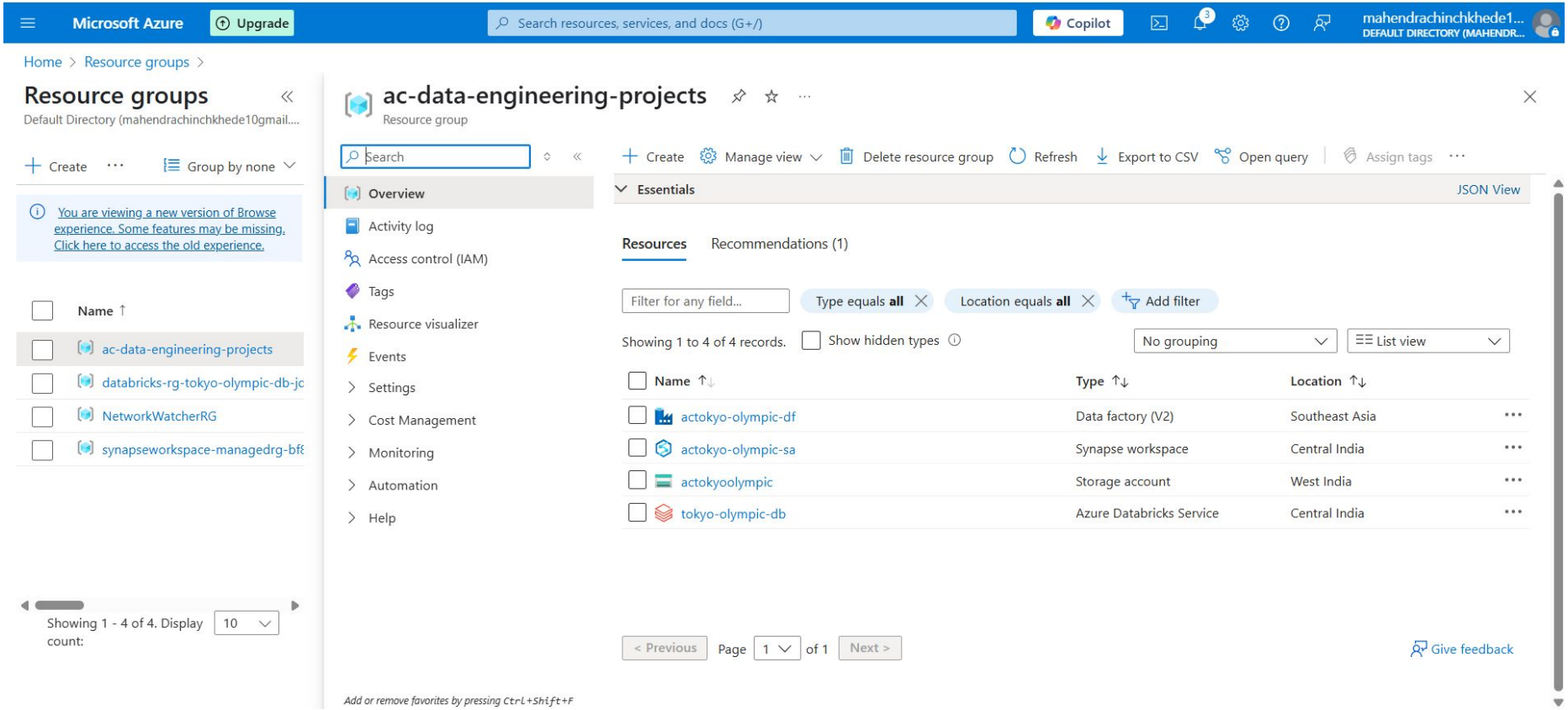
Calculate total medals won by each country



Calculate the average number of entries by gender for each discipline



PowerBI


Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
Copilot
mahendrachinchkhede1...
DEFAULT DIRECTORY (MAHENDR...
Home > Resource groups >
Resource groups
Default Directory (mahendrachinchkhede10gmail....
Create
Group by none
You are viewing a new version of Browse experience. Some features may be missing. Click here to access the old experience.
Name
ac-data-engineering-projects
databricks-rg-tokyo-olympic-db-jc
NetworkWatcherRG
synapseworkspace-managedrg-bf8
Showing 1 - 4 of 4. Display count:
10
ac-data-engineering-projects
Resource group
Search
Overview
Activity log
Access control (IAM)
Tags
Resource visualizer
Events
Settings
Cost Management
Monitoring
Automation
Help
Create
Manage view
Delete resource group
Refresh
Export to CSV
Open query
Assign tags
Essentials
JSON View
Resources
Recommendations (1)
Filter for any field...
Type equals all
Location equals all
Add filter
Showing 1 to 4 of 4 records.
Show hidden types
No grouping
List view
Name
Type
Location
actokyo-olympic-df
Data factory (V2)
Southeast Asia
actokyo-olympic-sa
Synapse workspace
Central India
actokyoolympic
Storage account
West India
tokyo-olympic-db
Azure Databricks Service
Central India
< Previous
Page 1 of 1
Next >
Give feedback
Add or remove favorites by pressing Ctrl+Shift+F