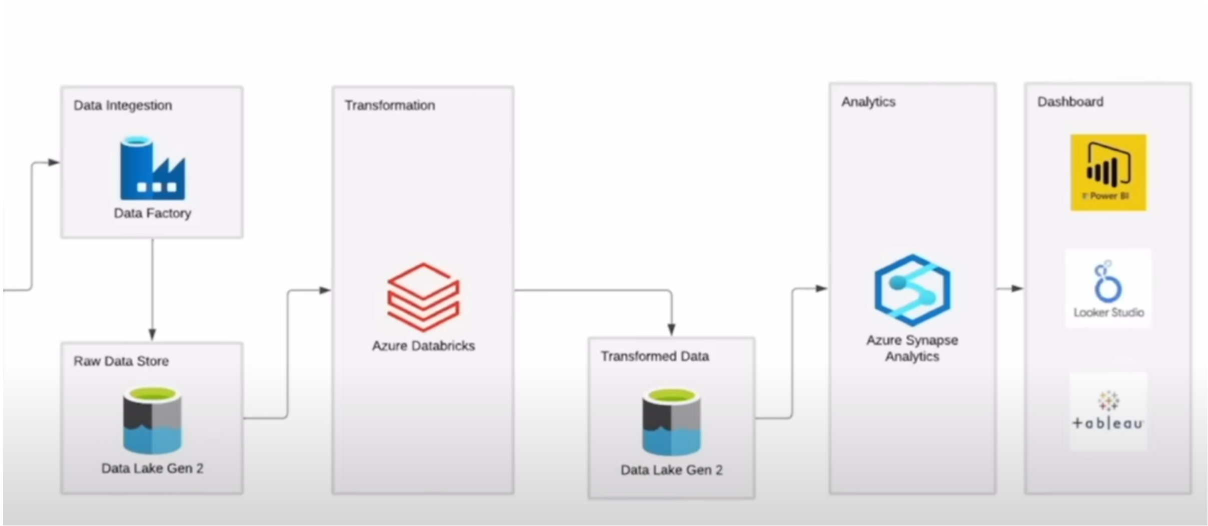
Dataflow diagram:



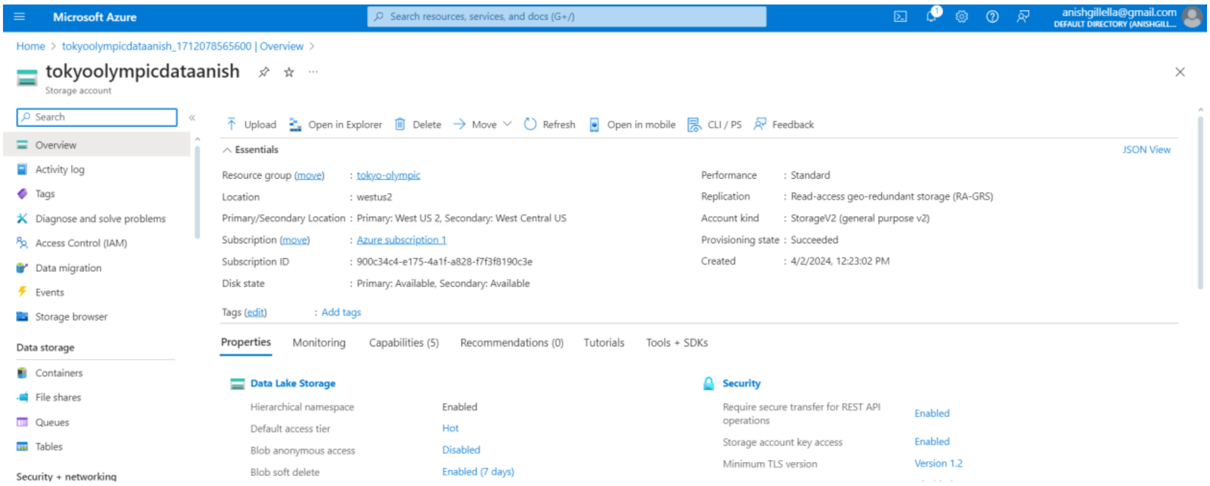
Data Factory – to create ETL pipelines to ingest the data and load the data into other locations

Data Lake Gen 2 – to combine data lake features with Azure Blob Storage

Databricks – Analytics platform built on top of Apache Spark for big data and ML frameworks

Synapse Analytics – Data warehouse available on Azure

Creating a storage account:



tokyoolympicdataanish

Storage account

Essentials

Resource group (move) : tokyo-olympic

Location : westus2

Primary/Secondary Location : Primary: West US 2, Secondary: West Central US

Subscription (move) : Azure subscription 1

Subscription ID : 900c34c4-e175-4a1f-a828-f7f3f8190c3e

Disk state : Primary: Available, Secondary: Available

Tags (edit) : Add tags

Performance : Standard

Replication : Read-access geo-redundant storage (RA-GRS)

Account kind : StorageV2 (general purpose v2)

Provisioning state : Succeeded

Created : 4/2/2024, 12:23:02 PM

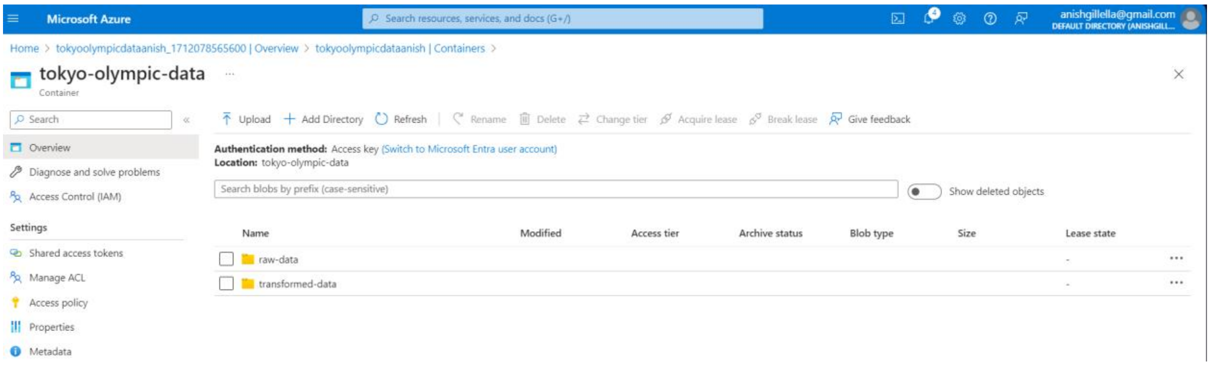
Data Lake Storage

Hierarchical namespace	Enabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)

Security

Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2

Storing the data in a container:



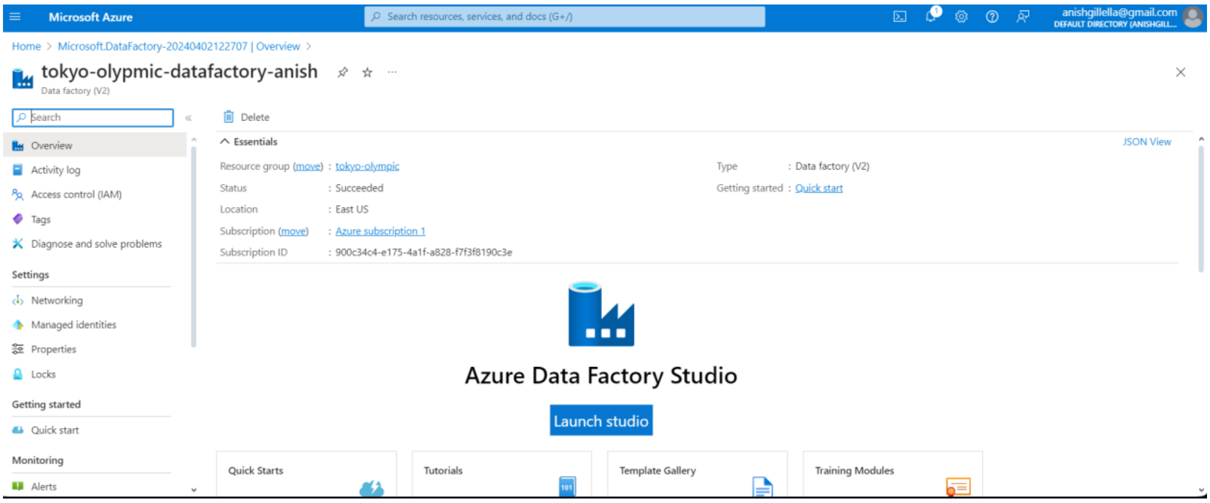
tokyo-olympic-data

Container

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: tokyo-olympic-data

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
□	raw-data						-	···
□	transformed-data						-	···

Creating a data factory:



tokyo-olypmic-datafactory-anish

Data factory (V2)

Resource group (move) : tokyo-olympic

Status : Succeeded

Location : East US

Subscription (move) : Azure subscription 1

Subscription ID : 900c34c4-e175-4a1f-a828-f7f3f8190c3e

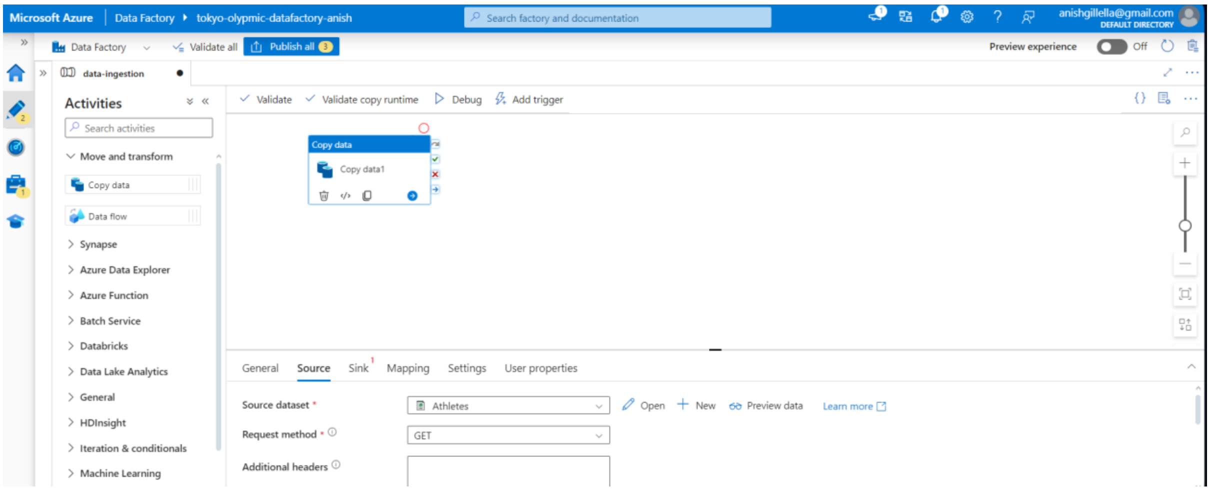
Type : Data factory (V2)

Getting started : Quick start

Azure Data Factory Studio

Launch studio

Forking the dataset from github and ingesting the data connecting the GitHub repository to the data factory:

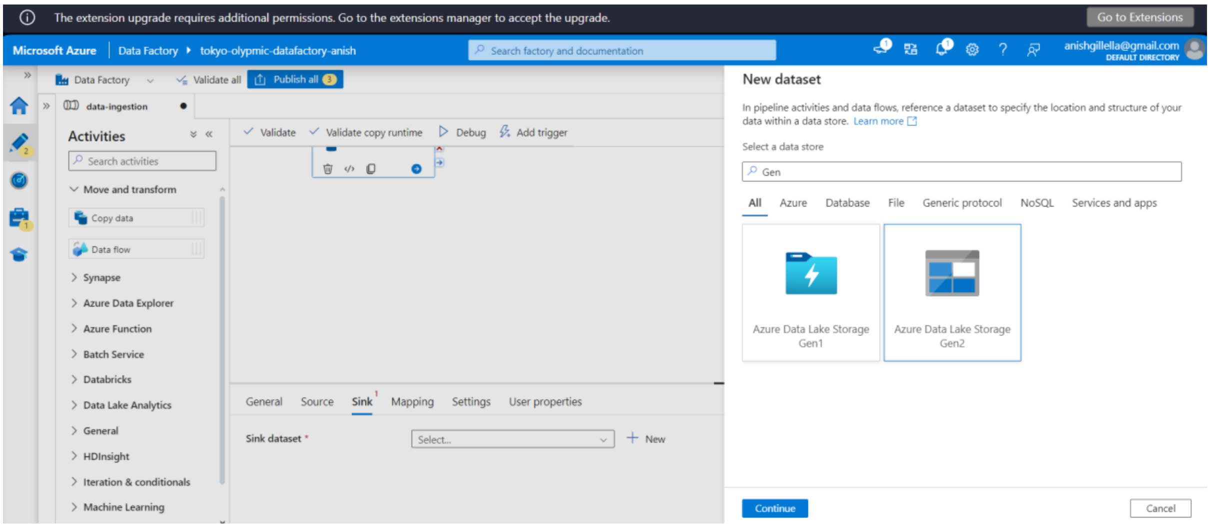


Source dataset * Athletes

Request method * GET

Additional headers

Sinking it into Gen 2 data lake storage:



The extension upgrade requires additional permissions. Go to the extensions manager to accept the upgrade.

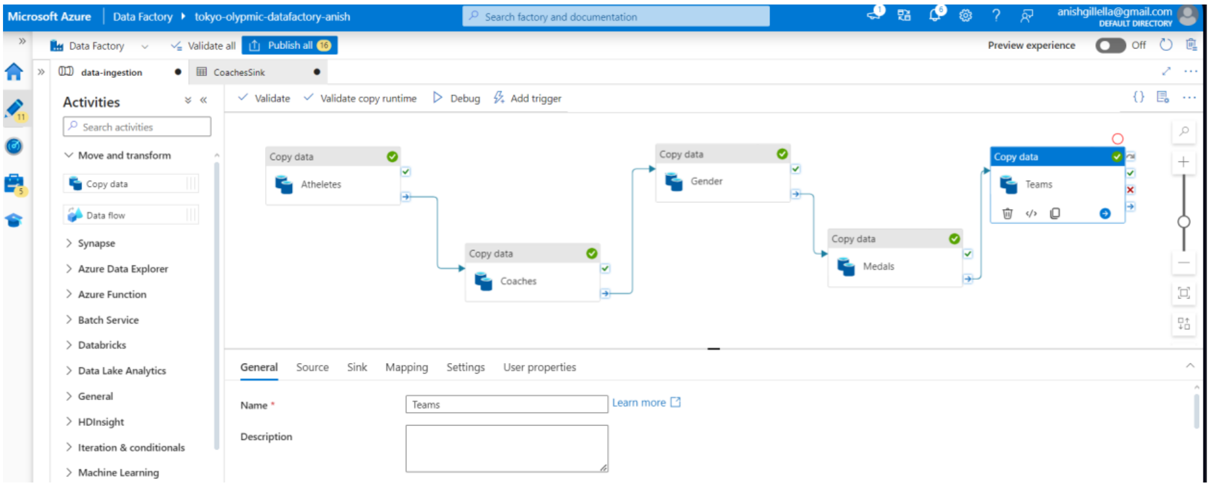
New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more

Select a data store: Gen

Sink dataset * Select...

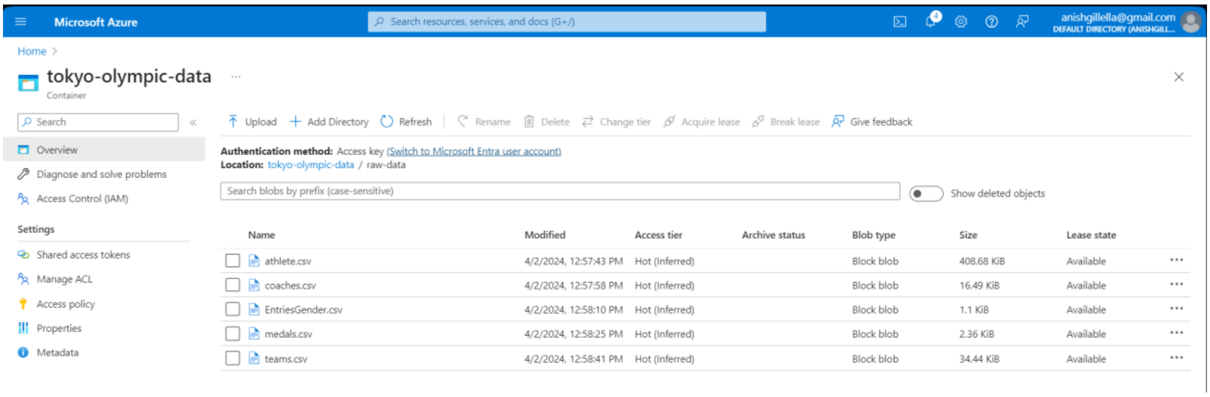
Bringing in all the data regarding the teams, medals, gender, coaches and athletes to the data factory:



Name * Teams

Description

Running the pipeline to import the data:

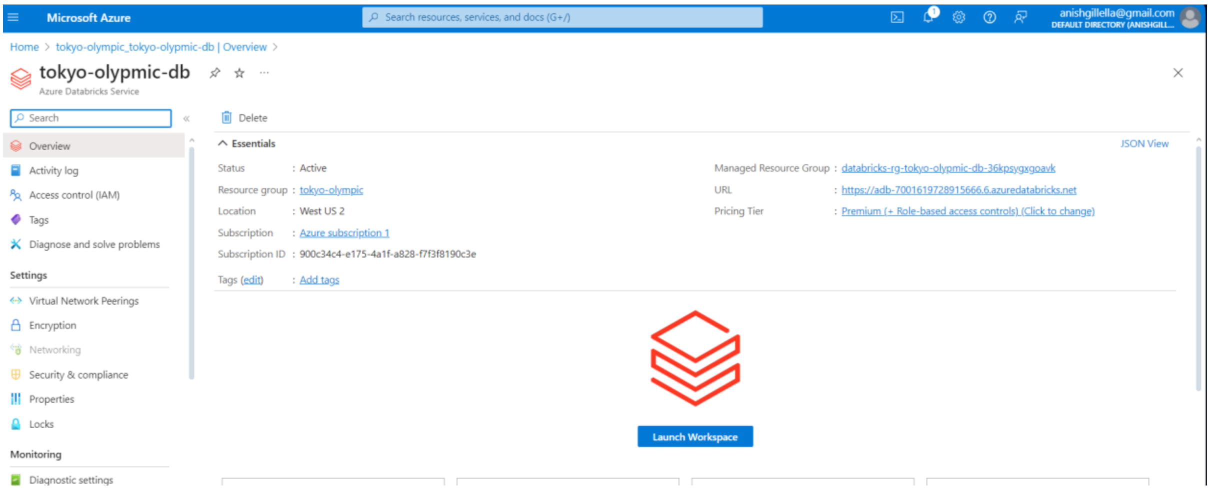


Authentication method: Access key (Switch to Microsoft Entra user account)
Location: tokyo-olympic-data / raw-data

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
athlete.csv	4/2/2024, 12:57:43 PM	Hot (Inferred)		Block blob	408.68 KiB	Available
coaches.csv	4/2/2024, 12:57:58 PM	Hot (Inferred)		Block blob	16.49 KiB	Available
EntriesGender.csv	4/2/2024, 12:58:10 PM	Hot (Inferred)		Block blob	1.1 KiB	Available
medals.csv	4/2/2024, 12:58:25 PM	Hot (Inferred)		Block blob	2.36 KiB	Available
teams.csv	4/2/2024, 12:58:41 PM	Hot (Inferred)		Block blob	34.44 KiB	Available

INGESTION AND STORING THE DATA HAS BEEN COMPLETED UPTO THIS POINT

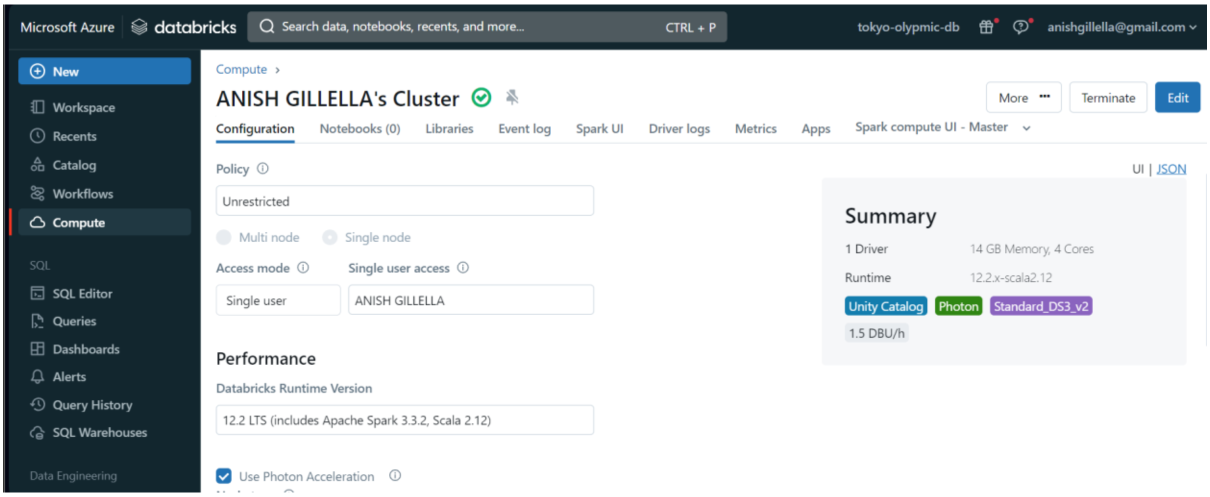
CREATING A DATABRICKS RESOURCE



Status : Active
Resource group : tokyo-olympic
Location : West US 2
Subscription : Azure subscription 1
Subscription ID : 900c34c4-e175-4a1f-a828-f7f3f8190c3e
Tags (edit) : Add tags

Managed Resource Group : databricks-rg-tokyo-olypmic-db-36kpsygxgoavk
URL : https://adb-7001619728915666.6.azuredatabricks.net
Pricing Tier : Premium (+ Role-based access controls) (Click to change)

CREATING A COMPUTE CLUSTER IN AZURE DATABRICKS



ANISH GILLELLA's Cluster

Policy: Unrestricted

○ Multi node ○ Single node

Access mode: Single user

Single user access: ANISH GILLELLA

Performance

Databricks Runtime Version: 12.2 LTS (includes Apache Spark 3.3.2, Scala 2.12)

☑ Use Photon Acceleration

Summary

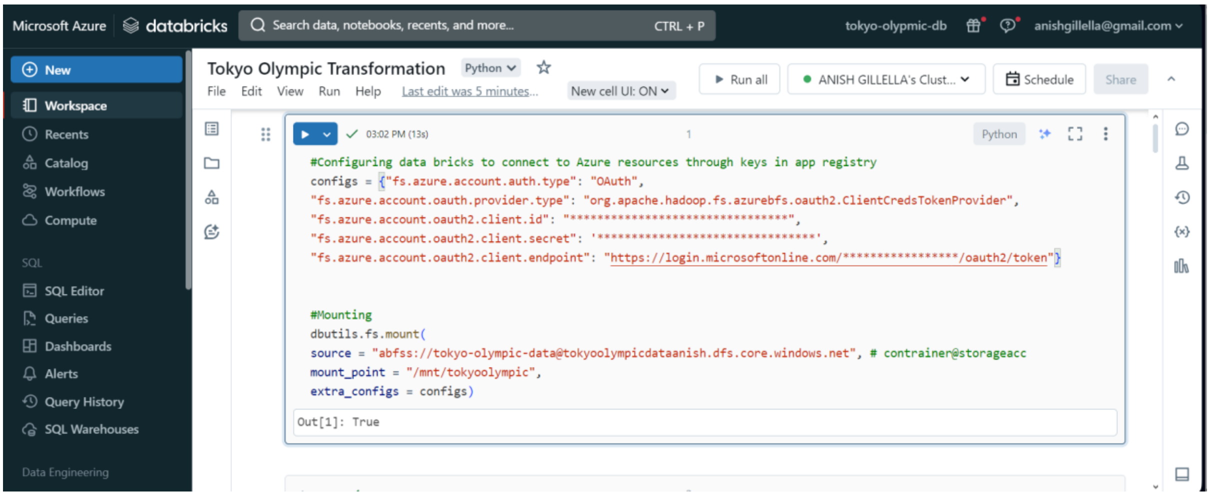
1 Driver	14 GB Memory, 4 Cores
Runtime	12.2.x-scala2.12

Unity Catalog Photon Standard_DS3_v2

1.5 DBU/h

Mounting azure data lake storage to the data bricks azure data factory to easy access the data

Configuring and Mounting the Notebook to the container in data storage account using the tenant ID and secret key:



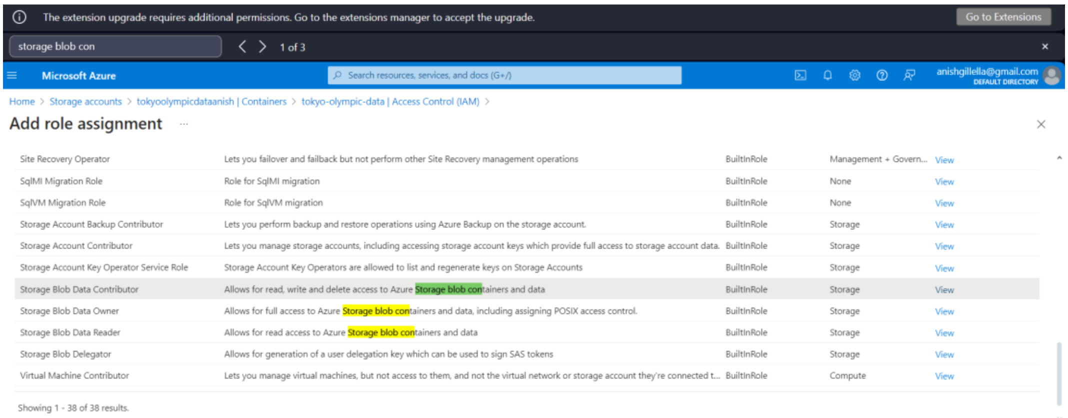
Tokyo Olympic Transformation

```
#Configuring data bricks to connect to Azure resources through keys in app registry
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": "********************************",
"fs.azure.account.oauth2.client.secret": '********************************',
"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/****************/oauth2/token"}


#Mounting
dbutils.fs.mount(
source = "abfss://tokyo-olympic-data@tokyoolympicdataanish.dfs.core.windows.net", # contrainer@storageacc
mount_point = "/mnt/tokyoolympic",
extra_configs = configs)
```

```
Out[1]: True
```

We do not have permission since the keys used in app registry do not have permission to access the contents of data storage account. So we assign the IAM in the containers to let the app registry access it:

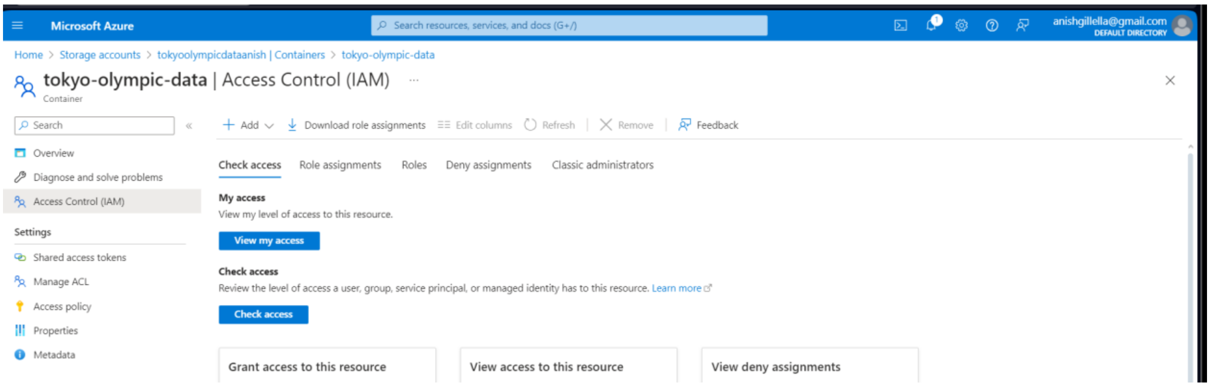

The extension upgrade requires additional permissions. Go to the extensions manager to accept the upgrade.
Go to Extensions
storage blob con
1 of 3
Microsoft Azure
Search resources, services, and docs (G+/)
anishgillella@gmail.com
DEFAULT DIRECTORY

Home › Storage accounts › tokyoolympicdataanish | Containers › tokyo-olympic-data | Access Control (IAM) ›

Add role assignment

Site Recovery Operator	Lets you failover and failback but not perform other Site Recovery management operations	BuiltInRole	Management + Govern...	View
SqlMI Migration Role	Role for SqlMI migration	BuiltInRole	None	View
SqlVM Migration Role	Role for SqlVM migration	BuiltInRole	None	View
Storage Account Backup Contributor	Lets you perform backup and restore operations using Azure Backup on the storage account.	BuiltInRole	Storage	View
Storage Account Contributor	Lets you manage storage accounts, including accessing storage account keys which provide full access to storage account data.	BuiltInRole	Storage	View
Storage Account Key Operator Service Role	Storage Account Key Operators are allowed to list and regenerate keys on Storage Accounts	BuiltInRole	Storage	View
Storage Blob Data Contributor	Allows for read, write and delete access to Azure Storage blob containers and data	BuiltInRole	Storage	View
Storage Blob Data Owner	Allows for full access to Azure Storage blob containers and data, including assigning POSIX access control.	BuiltInRole	Storage	View
Storage Blob Data Reader	Allows for read access to Azure Storage blob containers and data	BuiltInRole	Storage	View
Storage Blob Delegator	Allows for generation of a user delegation key which can be used to sign SAS tokens	BuiltInRole	Storage	View
Virtual Machine Contributor	Lets you manage virtual machines, but not access to them, and not the virtual network or storage account they're connected t...	BuiltInRole	Compute	View

Showing 1 - 38 of 38 results.



tokyo-olympic-data | Access Control (IAM)

Container

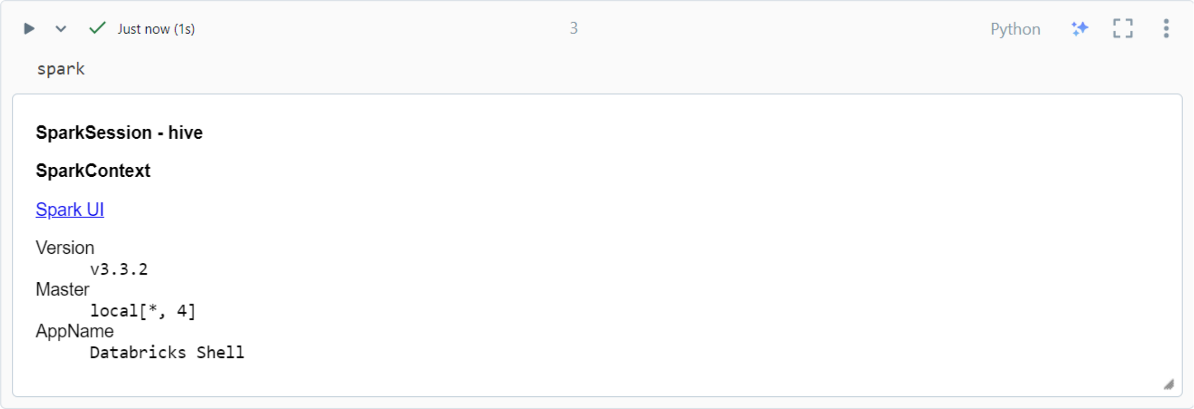
My access

View my level of access to this resource.

Check access

Review the level of access a user, group, service principal, or managed identity has to this resource. Learn more

Since it is azure databricks we do not have to create a spark session as it is already inbuilt like we usually do when using spark



```
spark
```

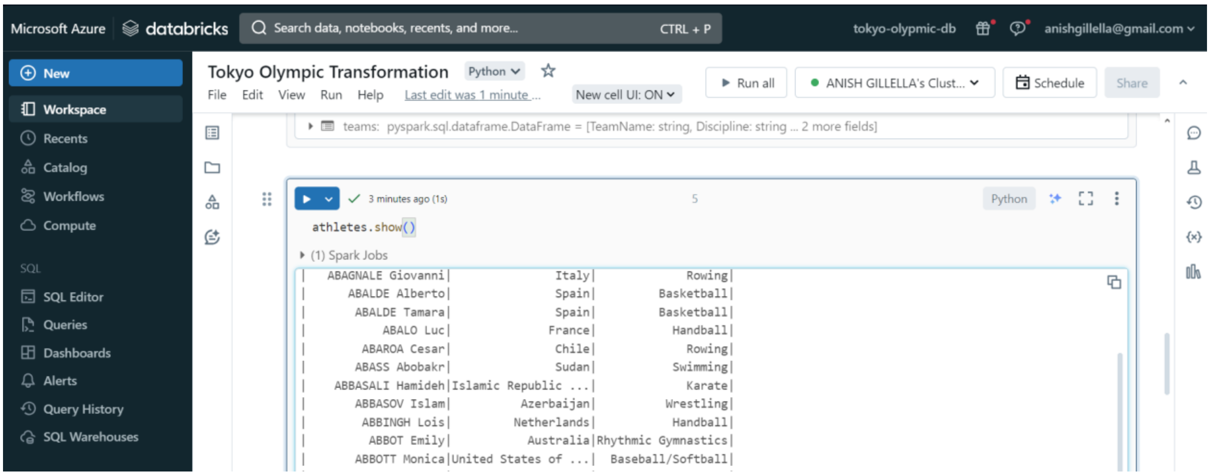
SparkSession - hive

SparkContext

Spark UI

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

Apache Spark uses lazy evaluation wherein spark does not perform any loading or transformation, it only performs the action when you call the particular data which was loaded with a function like we did using show() here

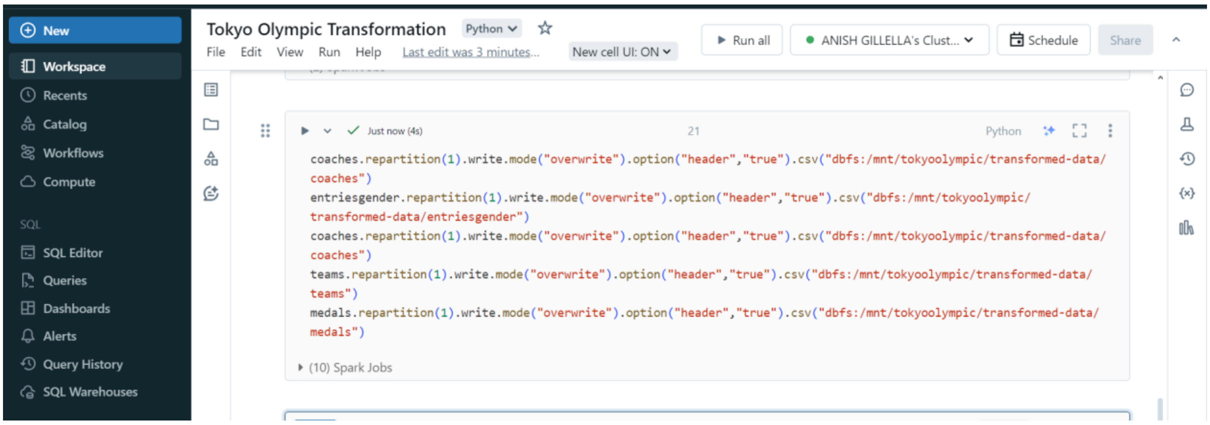


teams: pyspark.sql.dataframe.DataFrame = [TeamName: string, Discipline: string ... 2 more fields]

```
athletes.show()
```

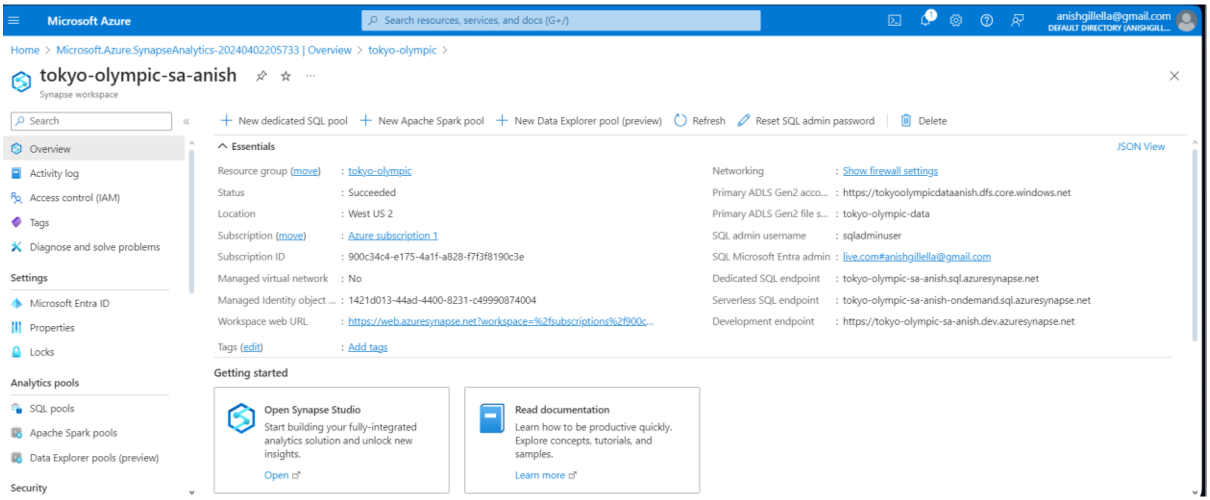
```
|   ABAGNALE Giovanni|               Italy|              Rowing|
|     ABALDE Alberto|               Spain|          Basketball|
|      ABALDE Tamara|               Spain|          Basketball|
|          ABALO Luc|              France|            Handball|
|       ABAROA Cesar|               Chile|              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|
```

Writing the transformed data back to the data storage factory



```
coaches.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolympic/transformed-data/
coaches")
entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolympic/
transformed-data/entriesgender")
coaches.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolympic/transformed-data/
coaches")
teams.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolympic/transformed-data/
teams")
medals.repartition(1).write.mode("overwrite").option("header","true").csv("dbfs:/mnt/tokyoolympic/transformed-data/
medals")
```

CREATING AN AZUR SYNAPSE ANALYTICS WORKSPACE



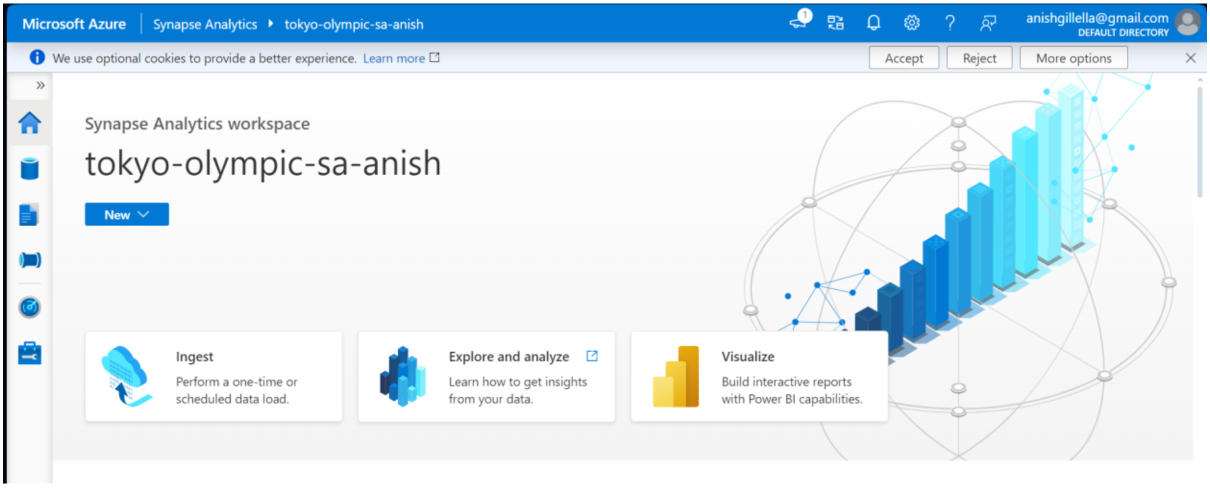
Resource group (move) : tokyo-olympic
Status : Succeeded
Location : West US 2
Subscription (move) : Azure subscription 1
Subscription ID : 900c34c4-e175-4a1f-a828-f7f3f8190c3e
Managed virtual network : No
Managed Identity object ... : 1421d013-44ad-4400-8231-c49990874004
Workspace web URL : https://web.azuresynapse.net?workspace=%2fsubscriptions%2f900c...
Tags (edit) : Add tags

Networking : Show firewall settings
Primary ADLS Gen2 acco... : https://tokyoolympicdataanish.dfs.core.windows.net
Primary ADLS Gen2 file s... : tokyo-olympic-data
SQL admin username : sqladminuser
SQL Microsoft Entra admin : live.com#anishgillella@gmail.com
Dedicated SQL endpoint : tokyo-olympic-sa-anish.sql.azuresynapse.net
Serverless SQL endpoint : tokyo-olympic-sa-anish-ondemand.sql.azuresynapse.net
Development endpoint : https://tokyo-olympic-sa-anish.dev.azuresynapse.net

Getting started

Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
Open

Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more



We use optional cookies to provide a better experience. Learn more

Synapse Analytics workspace

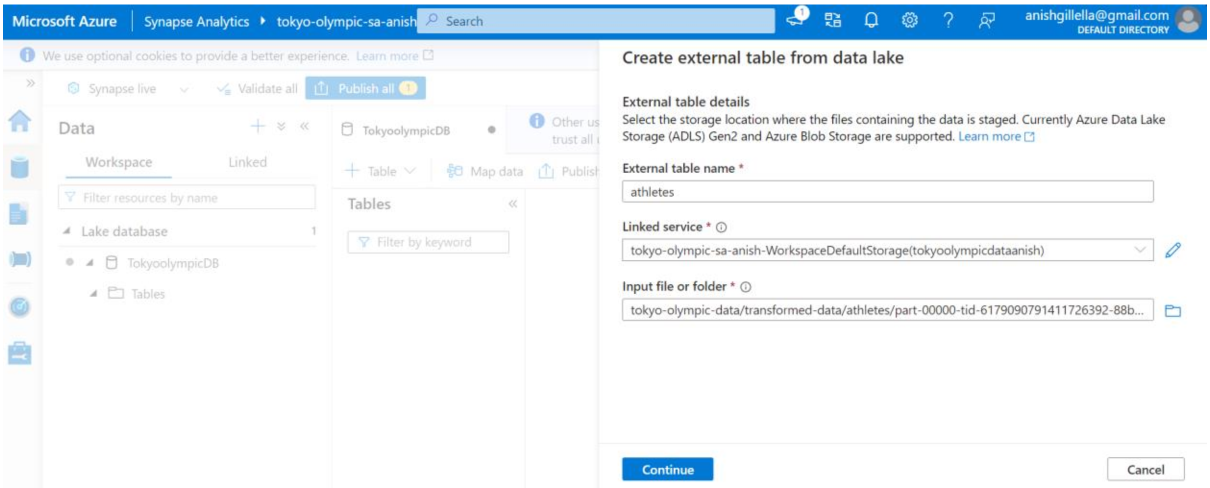
tokyo-olympic-sa-anish

Ingest
Perform a one-time or scheduled data load.

Explore and analyze
Learn how to get insights from your data.

Visualize
Build interactive reports with Power BI capabilities.

CREATING A TABLE IN SYNAPSE ANALYTICS FROM THE DATA LAKE



Create external table from data lake

External table details

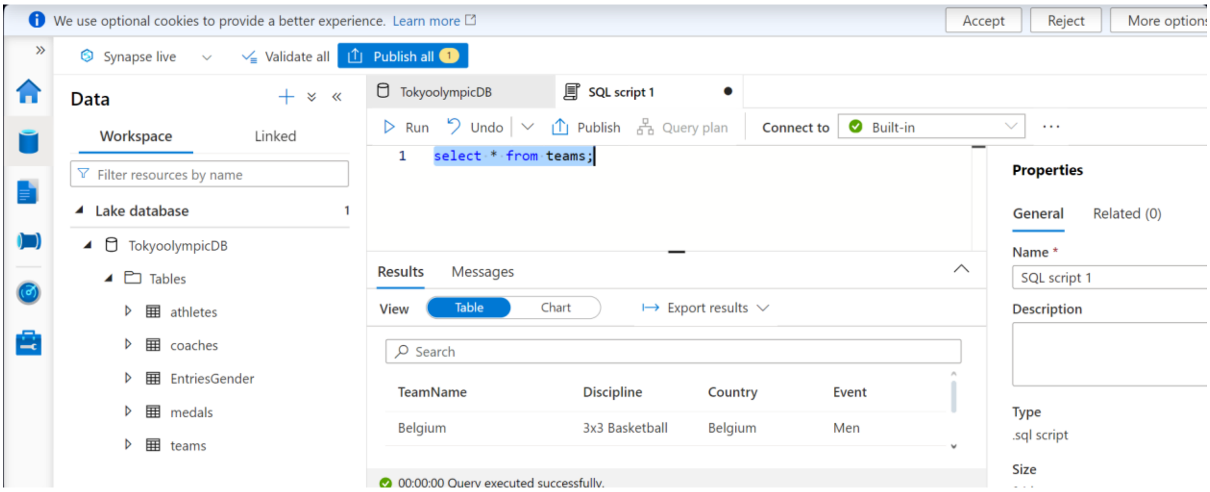
Select the storage location where the files containing the data is staged. Currently Azure Data Lake Storage (ADLS) Gen2 and Azure Blob Storage are supported. Learn more

External table name *: athletes

Linked service *: tokyo-olympic-sa-anish-WorkspaceDefaultStorage(tokyoolympicdataanish)

Input file or folder *: tokyo-olympic-data/transformed-data/athletes/part-00000-tid-6179090791411726392-88b...

QUERYING THE DATA IN SYNAPSE ANALYTICS:



```
select * from teams;
```

TeamName	Discipline	Country	Event
Belgium	3x3 Basketball	Belgium	Men

00:00:00 Query executed successfully.

Name *: SQL script 1

Description

Type: .sql script

Size