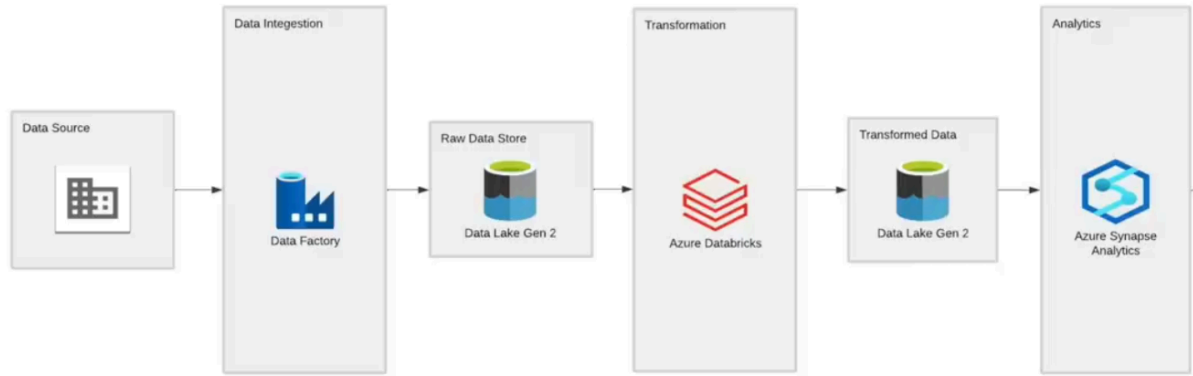


## Project 1 - Olympics Dataset Pipeline Creation

### Architecture Diagram:



### Olympic Dataset Overview

This contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics.

This dataset contains the details of the Athletes, Coaches, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches.

Source: Tokyo Olympics 2020 Website

### Five Files

- Athletes

#### Preview data

Linked service: AthletesHTTP

Object:

	PersonName	Country	Discipline
1	AALERUD Katrine	Norway	Cycling Road
2	ABAD Nestor	Spain	Artistic Gymnastics
3	ABAGNALE Giovanni	Italy	Rowing
4	ABALDE Alberto	Spain	Basketball
5	ABALDE Tamara	Spain	Basketball
6	ABALO Luc	France	Handball
7	ABAROA Cesar	Chile	Rowing
8	ABASS Abobakr	Sudan	Swimming
9	ABBASALI Hamideh	Islamic Republic of Iran	Karate
10	ABBASOV Islam	Azerbaijan	Wrestling

## - Coaches

Linked service: CoachesHTTP

Object:

	Name	Country	Discipline	Event
1	ABDELMAGID Wael	Egypt	Football	
2	ABE Junya	Japan	Volleyball	
3	ABE Katsuhiko	Japan	Basketball	
4	ADAMA Cherif	Côte d'Ivoire	Football	
5	AGEBA Yuya	Japan	Volleyball	
6	AIKMAN Siegfried Gottlieb	Japan	Hockey	Men
7	AL SAADI Kais	Germany	Hockey	Men
8	ALAMEDA Lonni	Canada	Baseball/Softball	Softball
9	ALEKNO Vladimir	Islamic Republic of Iran	Volleyball	Men
10	ALEKSEEV Alexey	ROC	Handball	Women

## - Entries Gender

Linked service: EntriesGenderHTTP

Object:

	Discipline	Female	Male	Total
1	3x3 Basketball	32	32	64
2	Archery	64	64	128
3	Artistic Gymnastics	98	98	196
4	Artistic Swimming	105	0	105
5	Athletics	969	1072	2041
6	Badminton	86	87	173
7	Baseball/Softball	90	144	234
8	Basketball	144	144	288
9	Beach Volleyball	48	48	96
10	Boxing	102	187	289

- Medals

Linked service: MedalsHTTP

Object:

	Rank	Team_Country	Gold	Silver	Bronze	Total	Rank by Total
1	1	United States of America	39	41	33	113	1
2	2	People's Republic of China	38	32	18	88	2
3	3	Japan	27	14	17	58	5
4	4	Great Britain	22	21	22	65	4
5	5	ROC	20	28	23	71	3
6	6	Australia	17	7	22	46	6
7	7	Netherlands	10	12	14	36	9
8	8	France	10	12	11	33	10
9	9	Germany	10	11	16	37	8
10	10	Italy	10	10	20	40	7

- Teams

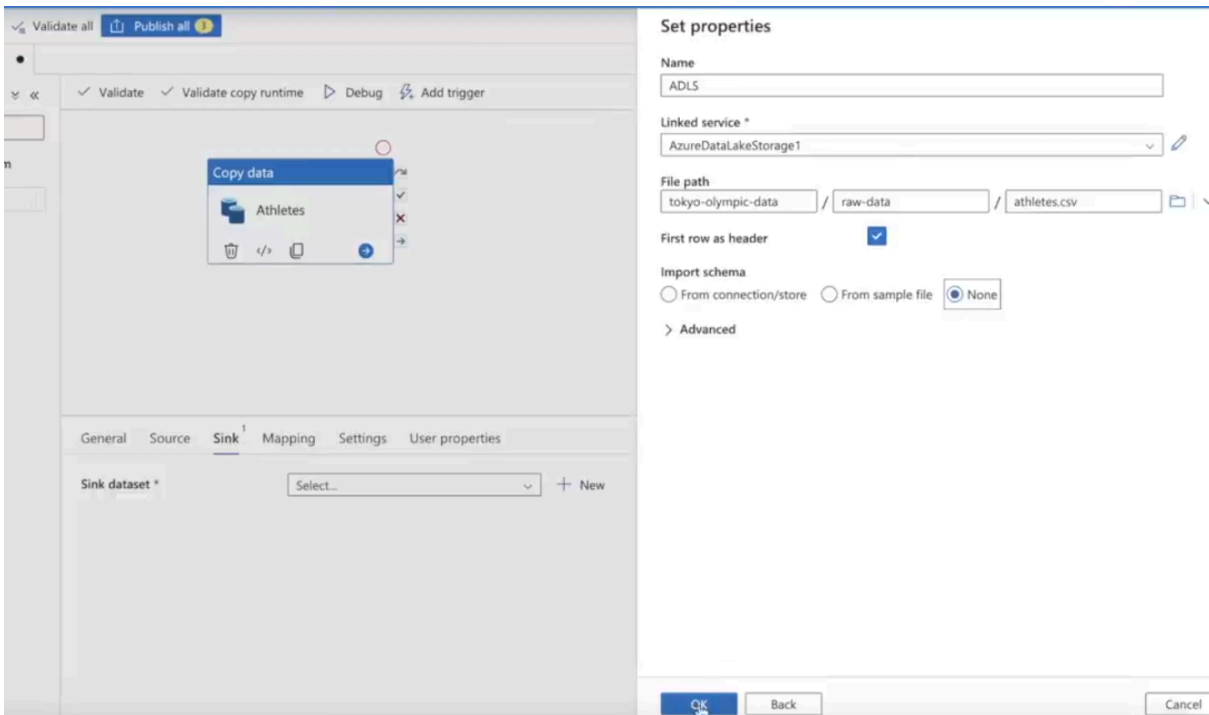
Linked service: TeamsHTTP

Object:

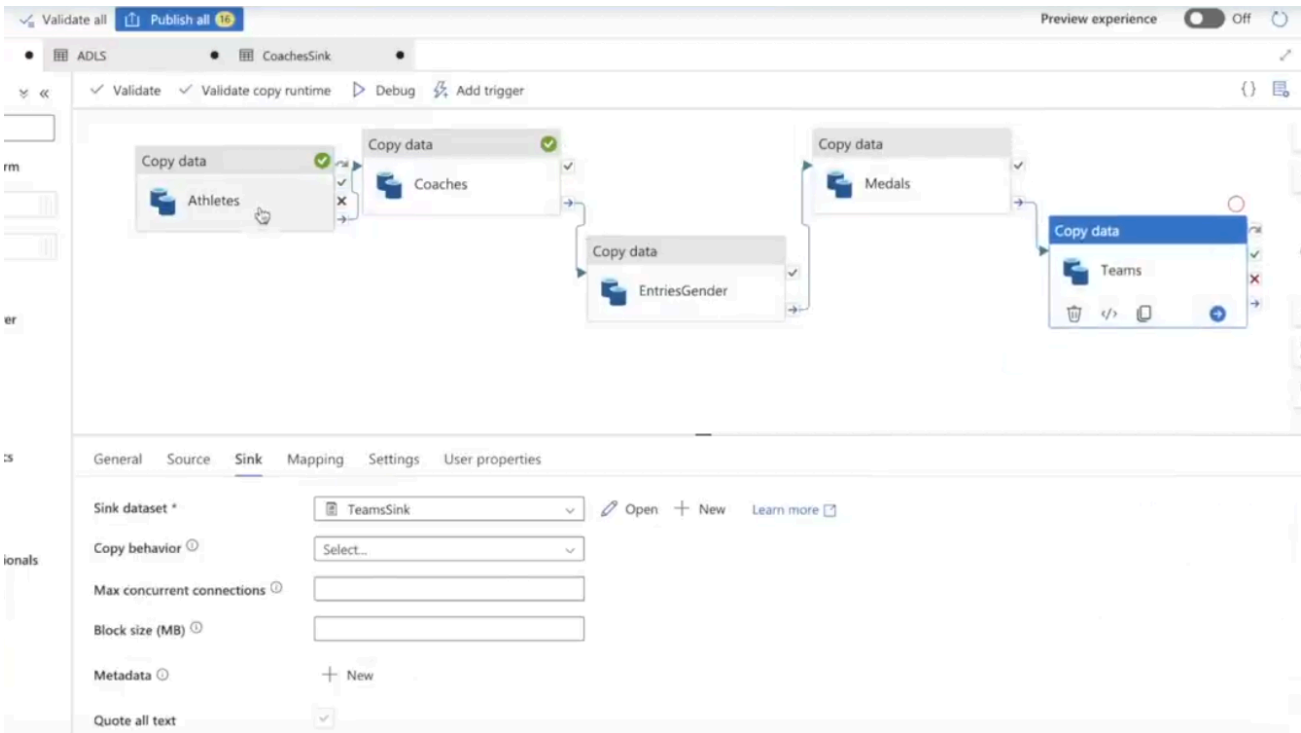
	TeamName	Discipline	Country	Event
1	Belgium	3x3 Basketball	Belgium	Men
2	China	3x3 Basketball	People's Republic of China	Men
3	China	3x3 Basketball	People's Republic of China	Women
4	France	3x3 Basketball	France	Women
5	Italy	3x3 Basketball	Italy	Women
6	Japan	3x3 Basketball	Japan	Men
7	Japan	3x3 Basketball	Japan	Women
8	Latvia	3x3 Basketball	Latvia	Men
9	Mongolia	3x3 Basketball	Mongolia	Women
10	Netherlands	3x3 Basketball	Netherlands	Men

Data Ingestion

Github HTTP Path to ADLS



Pipeline - copy data activity



## Raw-data container

Home > [tokyoolympicdata\\_1691561985395](#) | Overview > [tokyoolympicdata](#) | Containers >

**tokyo-olympic-data** Container

Search  Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: tokyo-olympic-data / raw-data

Search blobs by prefix (case-sensitive)  Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/> <a href="#">athletes.csv</a>		Hot (Inferred)		Block blob	408.68 KiB	Available
<input type="checkbox"/> <a href="#">coaches.csv</a>		Hot (Inferred)		Block blob	16.49 KiB	Available
<input type="checkbox"/> <a href="#">entriesgender.csv</a>		Hot (Inferred)		Block blob	1.1 KiB	Available
<input type="checkbox"/> <a href="#">medals.csv</a>		Hot (Inferred)		Block blob	2.36 KiB	Available
<input type="checkbox"/> <a href="#">teams.csv</a>		Hot (Inferred)		Block blob	34.44 KiB	Available

## Data Transformation - csv column data types transformed

Home > [tokyo-olympic\\_tokyo-olympic-db](#) | Overview >

**tokyo-olympic-db** Azure Databricks Service

Search  Delete

Overview

Activity log

Access control (IAM)

Tags

Settings

Virtual Network Peering

Encryption

Networking

Properties

Locks

Monitoring

Dagnostic settings

Essentials

Status : Active

Resource group : [tokyo-olympic](#)

Location : Southeast Asia

Subscription : [Free Trial](#)


Subscription ID : 1d0b10b3-5f84-4670-a1aa-e0c6b631e878

Tags (edit) : [Add tags](#)

Managed Resource Group : [databricks-rg-tokyo-olympic-db-ivykgign2k7ii](#)

URL : <https://adb-2723700013584314.14.azuredatabricks.net>

Pricing Tier : [Premium \(+ Role-based access controls\)](#) (Click to change)



[Launch Workspace](#)

## App registration - to make connection between Azure Databricks and ADLS

Home > App registrations >

**app01**

Search  Delete Endpoints Preview features

Overview

Quickstart

Integration assistant

Manage

Branding & properties

Authentication

Certificates & secrets

Token configuration

API permissions

Expose an API

App roles

Owners

Create application

Successfully created application app01.

Got a second? We would love your feedback on Microsoft identity platform (previously Azure AD for developer). →

Essentials

Display name : [app01](#)

Application (client) ID : [a43aff44-6236-4764-8f52-1a3f2dc193f9](#)

Object ID : [f00d5f2b-ad48-459b-82ab-26f131859699](#)

Directory (tenant) ID : [86a0360b-c00f-46b5-a3a7-6919a683e3a3](#)

Supported account types : [My organization only](#)

Client credentials : [Add a certificate or secret](#)

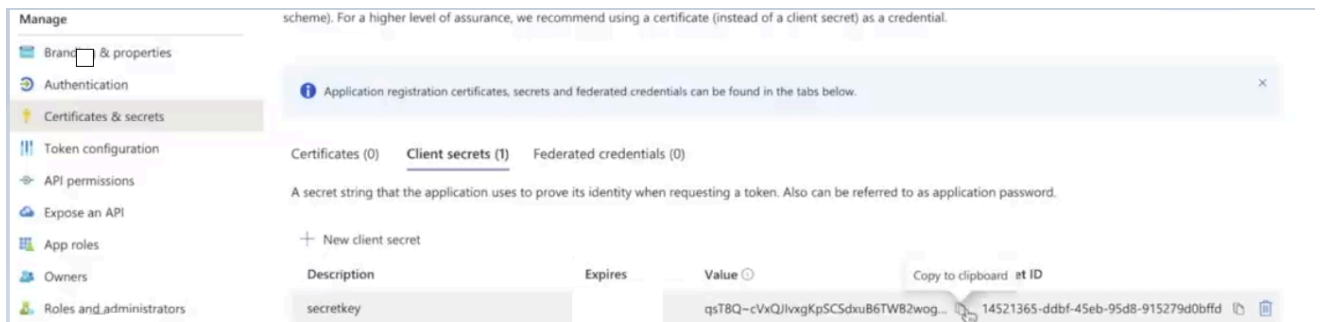
Redirect URIs : [Add a Redirect URI](#)

Application ID URI : [Add an Application ID URI](#)

Managed application in L... : [app01](#)

Welcome to the new and improved App registrations. Looking to learn how it's changed from App registrations (Legacy)? [Learn more](#)

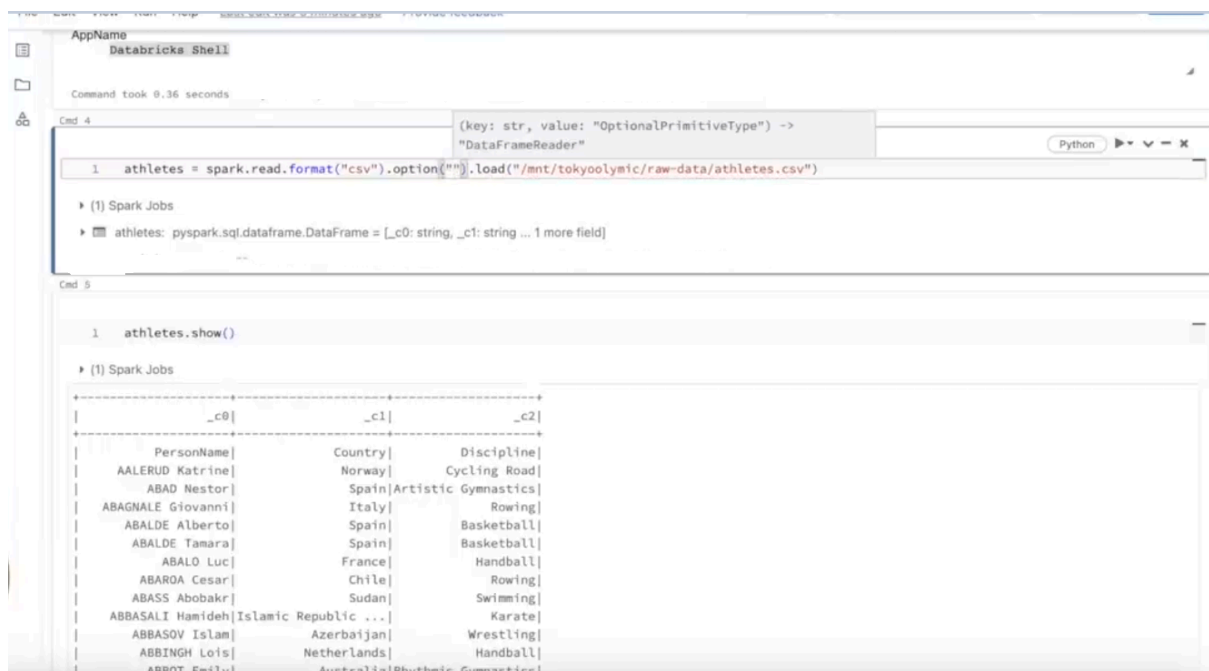
Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure AD Graph. We will continue to provide technical support and security updates but we will no longer provide feature updates. Applications will need to be upgraded to Microsoft Authentication Library (MSAL) and Microsoft Graph. [Learn more](#)



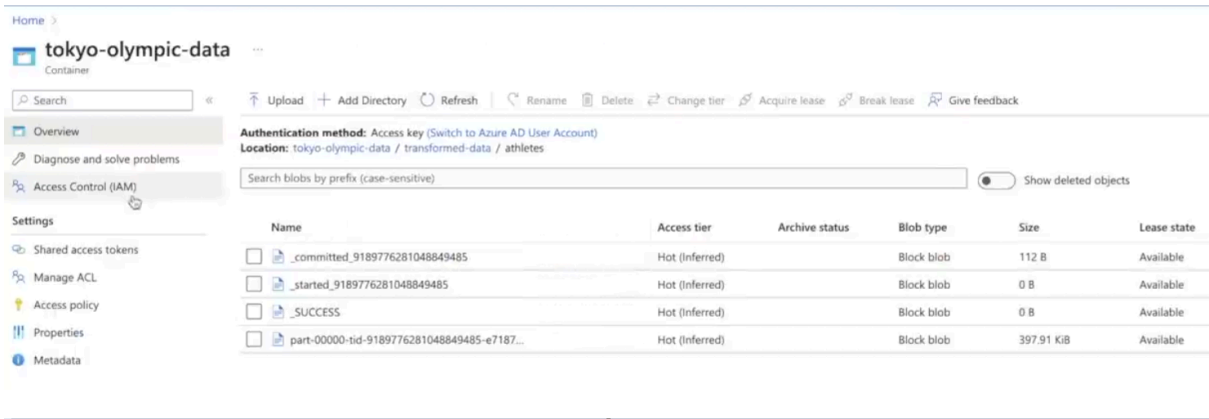
## # Configuration Settings

```
configs= {"fs.azure.account.auth.type": "OAuth",
"fs.azure.account.oauth.provider.type":
"org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsToken Provider",
"fs.azure.account.oauth2.client.id":
"a43aff44-6236-4764-8f52-1a3f2dc193f9",
"fs.azure.account.oauth2.client.secret":
'qsT8Q~cVxQJlvxgKpSCSdxuB6TWB2wogG9661dlu',
"fs.azure.account.oauth2.client.endpoint":
"https://login.microsoftonline.com/86a0360b-c00f-46b5-a3a7-6919a683e3a3/
oauth2/token"}
```

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

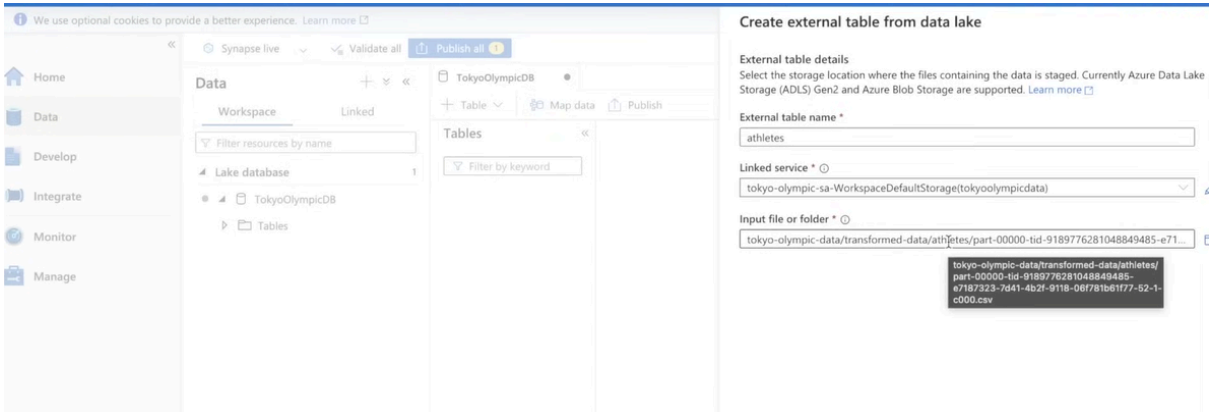


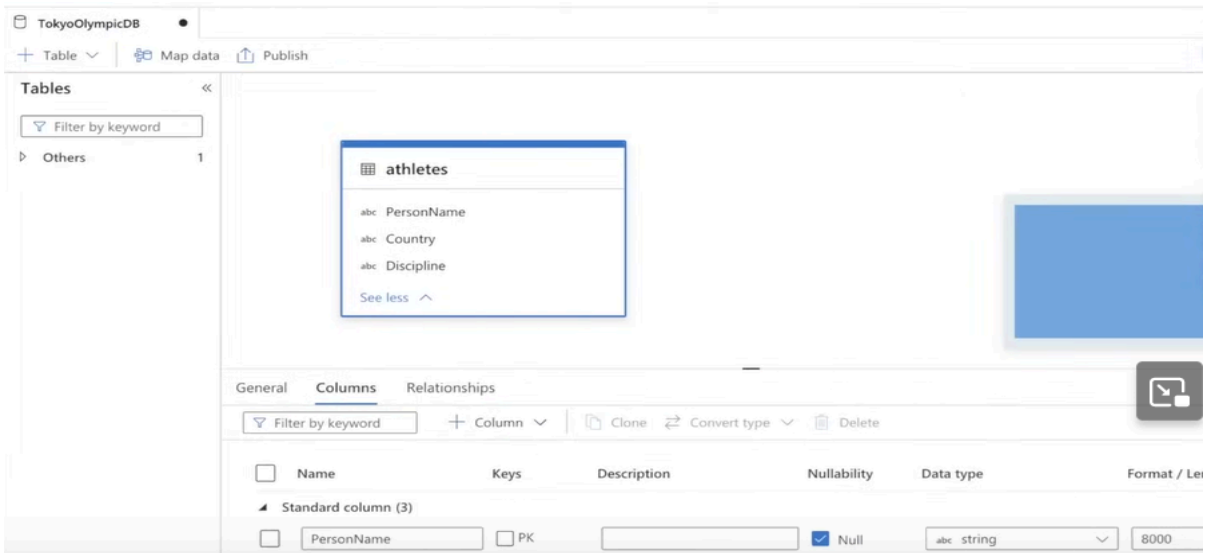
```
# Load the data in processed container
athletes.repartition(1).write.mode("overwrite").option("header", 'true').
csv("/mnt/tokyoolympic/transformed-data/athletes")
coaches.repartition(1).write.mode("overwrite").option("header", "true").c
sv("/mnt/tokyoolympic/transformed-data/coaches")
entriesgender.repartition(1).write.mode("overwrite").option("header", "tr
ue").csv("/mnt/tokyoolympic/transformed-data/entriesgender")
medals.repartition(1).write.mode("overwrite").option("header", "true").cs
v("/mnt/tokyoolympic/transformed-data/medals")
teams.repartition(1).write.mode("overwrite").option("header", "true").csv
("/mnt/tokyoolympic/transformed-data/teams")
```



Data Loading

Create Lake DB





Copy all tables in lake db

