**Azure project: Building Data Pipelines in Azure with Azure Synapse Analytics**

**Summary:** In this project, we will build a pipeline in Azure using Azure Synapse Analytics, Azure Storage and Azure Synapse SQL pool to perform data analysis on the 2021 Olympics dataset.

**Software:** Azure Storage Account, Azure Synapse Analytics

**Contents**

[**Instructions** 1](#_Toc172575017)

[1. Dataset 1](#_Toc172575018)

[2. Project Architecture 2](#_Toc172575019)

[3. Project Workflow 2](#_Toc172575020)

[a. Create an azure storage account and upload data files in a container 2](#_Toc172575021)

[b. Create an azure synapse analytics workspace. 2](#_Toc172575022)

[c. Create a SQL pool in azure synapse workspace. 2](#_Toc172575023)

[d. Create table structure in SQL pool. 3](#_Toc172575024)

[e. Create an ADF data pipeline to ingest data from azure storage into SQL pool tables. 5](#_Toc172575025)

# **Instructions**

## Dataset

The Tokyo 2021 Olympics dataset refers to the collection of data related to the Summer Olympic Games that took place in Tokyo, Japan in 2021.

* **Athletes file (PersonName, Country, Discipline)**
* **Coaches file (Name, Country, Discipline)**
* **EntriesGender file (Discipline, Female, Male, Total)**
* **Medals file (Rank, Team\_country, Gold, Silver, Bronze, Total, Rank by total)**
* **Teams file (Teamname, Discipline, Country, Event)**

## Project Architecture

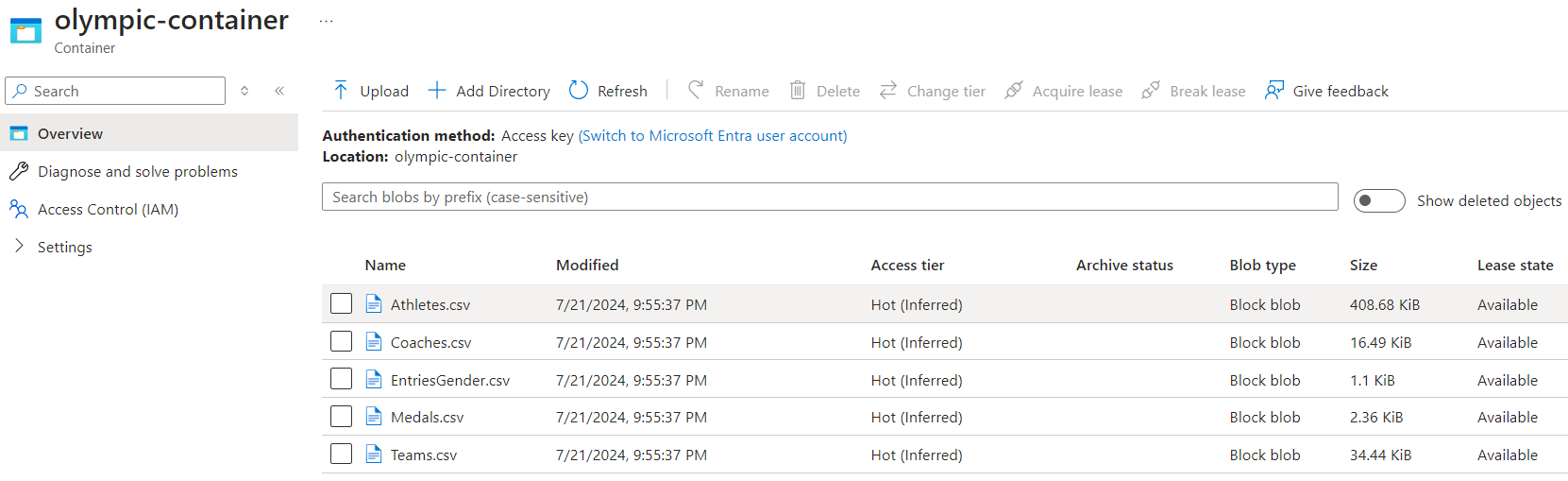
A diagram of a data factory

Description automatically generated

## Project Workflow

### Create an azure storage account and upload data files in a container

* Create an Azure Storage account:
  + Go to the Azure Portal, sign in > create Storage account > Fill in the required fields for the storage account, such as the name, subscription, resource group, and location.
  + Choose "Blob storage" as the account kind, appropriate redundancy options > Click the "Review + create"
* Create a container: **olympic-container**
  + Go to your newly created Storage account > Click on the "Containers" & create a new container > Give container a name, and choose the appropriate access level > Click the "Create".
  + Upload a CSV file



### Create an azure synapse analytics workspace.

* Create "Azure Synapse Workspace" > fill in the required information, such as the workspace name, subscription, resource group, and location > Choose the storage account type and configure the firewall settings, if desired > Click on the "Review + create"

### Create a SQL pool in azure synapse workspace.

Create a Dedicated SQL Pool (in Overview)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### Create table structure in SQL pool.

Open the Synapse studio > Go to Data, Workspace, SQL database, and open the newly created dedicated SQL pool > Click on three dots in front of the Tables, and create a new SQL script.

A screenshot of a computer

Description automatically generated

Add the following DDL scripts to create 5 tables for each data file > Run

|  |
| --- |
| **/\*\*\*\*\*\*Object: Table [dbo].[AthletesOlympics] \*\*\*\*\*\*/**  **SET ANSI\_NULLS ON**  **GO**  **SET QUOTED\_IDENTIFIER ON**  **GO**  **CREATE TABLE [dbo].[AthletesOlympics](**  **[PersonName] [nvarchar](100) NULL,**  **[Country] [nvarchar](100) NULL,**  **[Discipline] [nvarchar](100) NULL**  **)**  **GO**  **/\*\*\*\*\*\* Object: Table [dbo].[CoachesOlympics] \*\*\*\*\*\*/**  **SET ANSI\_NULLS ON**  **GO**  **SET QUOTED\_IDENTIFIER ON**  **GO**  **CREATE TABLE [dbo].[CoachesOlympics](**  **[CoachName] [nvarchar](100) NULL,**  **[Country] [nvarchar](100) NULL,**  **[Discipline] [nvarchar](100) NULL,**  **[Event] [nvarchar](50) NULL**  **)**  **GO**  **/\*\*\*\*\*\* Object: Table [dbo].[AthletesGenderOlympics] \*\*\*\*\*\*/**  **SET ANSI\_NULLS ON**  **GO**  **SET QUOTED\_IDENTIFIER ON**  **GO**  **CREATE TABLE [dbo].[AthletesGenderOlympics](**  **[Discipline] [nvarchar](100) NULL,**  **[Male] [int] NULL,**  **[Female] [int] NULL,**  **[TotalAthletes] [int] NULL**  **)**  **GO**  **/\*\*\*\*\*\* Object: Table [dbo].[CountryMedalsOlympics] \*\*\*\*\*\*/**  **SET ANSI\_NULLS ON**  **GO**  **SET QUOTED\_IDENTIFIER ON**  **GO**  **CREATE TABLE [dbo].[CountryMedalsOlympics](**  **[RankId] [int] NULL,**  **[Country] [nvarchar](100) NULL,**  **[Gold] [int] NULL,**  **[Silver] [int] NULL,**  **[Bronze] [int] NULL,**  **[Total] [int] NULL,**  **[RankByTotal] [int] NULL**  **)**  **GO**  **/\*\*\*\*\*\* Object: Table [dbo].[TeamOlympics] \*\*\*\*\*\*/**  **SET ANSI\_NULLS ON**  **GO**  **SET QUOTED\_IDENTIFIER ON**  **GO**  **CREATE TABLE [dbo].[TeamOlympics](**  **[TeamName] [nvarchar](100) NULL,**  **[Country] [nvarchar](100) NULL,**  **[Discipline] [nvarchar](100) NULL,**  **[Event] [nvarchar](50) NULL**  **)**  **GO** |

It will show 5 tables

A screenshot of a computer

Description automatically generated

### Create an ADF data pipeline to ingest data from azure storage into SQL pool tables.

Go to your Synapse workspace > open Synapse Studio > click "Integrate" > create a new ADF pipeline

A screenshot of a computer

Description automatically generated

Use the “Copy data” component to move and transform data from one location to another > Define source and sink

* For the **source** > Select Azure Data Lake Storage Gen2 > Select DelimitedText format > create a Linked service for Azure Data Lake Storage (choose New if not created yet + provide storage account name)
* In Set properties console, input file path stored in storage account (or containers)

A screenshot of a computer

Description automatically generated

For the sink > Click on New to create a Sink dataset > Select Azure Synapse dedicated SQL Pool > Select the SQL pool and the table where you want to ingest the data.

Over here, we are trying to ingest “Athletes.csv” file, hence we select the “AthletesOlympics” table

A screenshot of a computer

Description automatically generatedRemember in Sink, choose Bulk insert as Copy method > inserting large amounts of data into a database table using a single SQL command.

A screenshot of a computer

Description automatically generated

Next, go to Mapping section (f defining the relationships between the source and target data in a data integration scenario) > Click on Import schemas > it will automatically map the columns of the source dataset and sink dataset  
A screenshot of a computer

Description automatically generated

Then, click Validate to check if there is any error > If no error, To run the pipeline, click on Debug. In Output, it will show status as Succeeded.

A screenshot of a computer

Description automatically generated

We have successfully copied the data of ‘Athletes.csv’ file stored in Azure storage into the ‘AthletesOlympics’ table stored in dedicated SQL Pool. Please follow the same steps for the remaining data files.

Check 100 record of file AthletesOlympic, it shows data has been loaded into already.

A screenshot of a computer

Description automatically generated