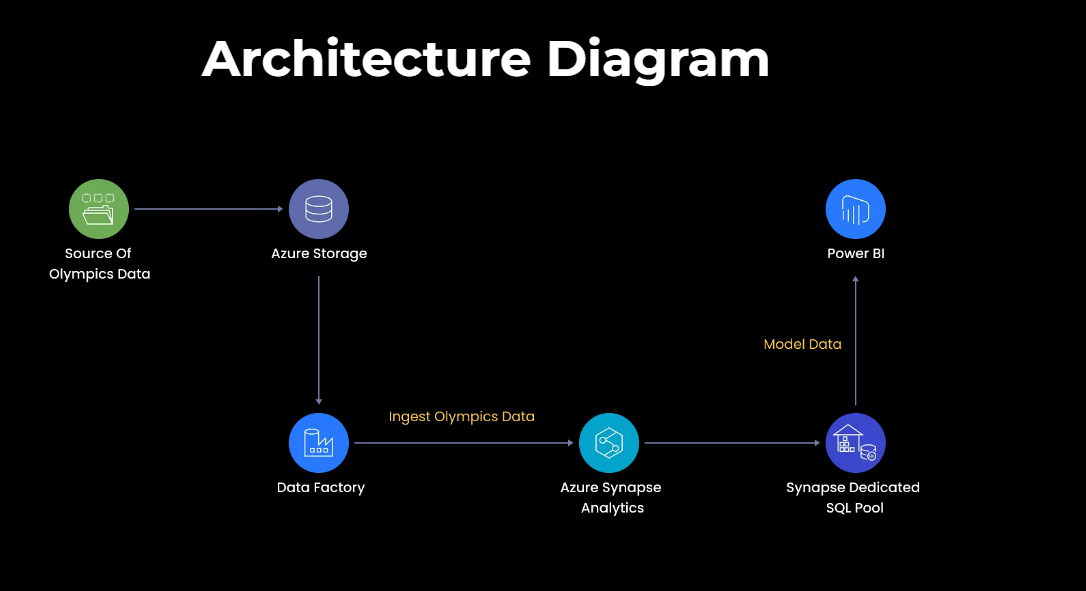
**Tokyo Olympic Pipeline Project:**

Project Architecture:



The Source of the Olympic data is Kaggle and stored into Github repositories and from github to Azure Data Lake Storage account Gen2.

Then the data will be moved to the object container (Data Lake) through Azure data factory with the help of copy activity.

After data is being stored in Azure data factory, We transfer the data to the Azure Databricks through the connection.

Once data is modified in Azure Databricks, we store that data into Azure data Lake container with the help of pyspark code.

After Transformation we take that data into Azure Synapse Analytics tool for data visualization and reporting through SQL commands, we perform transformation.

We can transfer the data into Power BI or Tableau for the reporting.

A screenshot of a computer

Description automatically generated

>>Create a Storage Gen2 Data Lake account.

>> Create a folder in container with name: Tokyo-Olympic-data.

>>Create 2 folders inside Tokyo-Olympic-data Folder for Source data and transformed data.

>>Create Azure Data Factory account.

A screenshot of a computer

Description automatically generated

>>Go to Activity.

>>Drag and Drop Copy Activity.

>>Go to Source << Source Dataset << + New << Select Source Data as HTTP.

>>Select File format as Delimited Text (CSV).

>>Add the Source and Sink.

>>Preview the data after Linking and Sinking the data.

>>Validate and debug the Activity.

>>Copy the copy activity and change the Source and destination for the other files.

>>After loading the data into Datalake Gen2, Create a Databricks account with the default setting.

>>Open Databricks and create a compute cluster with the default settings.

A screenshot of a computer

Description automatically generated

>>Open a new workstation and rename the workstation name and connect the cluster which was created on Compute.

A screenshot of a computer

Description automatically generated

>>To connect Azure storage with databricks, go to Azure search and open app registration.

>>Enter the app name and create the app.

>>Copy the Application(ClientID) and TenantID from the app which is created.

Application (client) ID: 676bd3b6-2726-4bea-b13d-aef643517354

Directory (tenant) ID: e24ac094-efd8-4a6b-98d5-a129b32a8c9a

>>open certificates and Secrects from the App and Click on New Client Secret.

>>Create a SecretID for Accessing the Azure datalake.

>>Copy the Secret key Value from the description: 698f565f-caf3-4ad1-87f9-dc8526fdfbdc

>>Paste all the credentials taken from the app in the configuration code in workstation.

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/Paste\_**tanent\_id**/oauth2/token"}

>>copy the mount code and paste the folder and file names in the mount code.

dbutils.fs.mount(

source = "abfss://PasteContainerName@StorageAccount.dfs.core.windows.net", # contrainer@storageacc

mount\_point = "/mnt/tokyoolymic",

extra\_configs = configs)

>>Source link: Data = paste the container name. After @ = Paste storage name.

>>After configuration run the code.

>>If connection is successful then use %fs ls “/mnt/tokyoolympic”.

>>If mount code is throwing error then, go to storage>>container>>Access IAM>>+Add>>Search Storage Blob Data Contributor and select>>Next>>Click on Members and search the app name which was created for security and select it. >>Next >>Review and Assign.(This process is for giving the access to the app to give the read write and delete permission or any object from the container).

>>Run the %fs code.

>>Once connection is done. Load the files to the workstation and start transformation.

>> After transformation paste the data into Azure Data Lake Gen2 container which was created for transformed data.

>>The following code to transfer data from adf to the container.

#To store the transformed data into Azure datalake:

##.mode("overwrite") is for overwrtitng the file in the folder.

##.repartition(1) is to partition the into specified value.

athletes.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/athletes")

coaches.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/coaches")

entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/entriesgender")

medals.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/medals")

teams.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymic/transformed-data/teams")

>>Once data is transferred to the container, Create a synapse account.

>>Open synapse analytics, open Data section >> +Linked >>select lake database. Select the folders from data lake container.

>>Name the database name and click on >> +Table >> Select From data lake >> Name the table name and connect it with the required db.

>>Once data is loaded to Synapse analytics, we can validate and publish the data.

A computer screen shot of a computer

Description automatically generated

>>After publishing the data, we can perform SQL operations based on our requirements.