Walk Through steps for Data Tutorials, Azure End-to-End (Data Analysis) by Yeo Chee En Luke

NOTE:

1. Where to get the data:

[https://www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo](%20https:/www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo)

1. For this project , I will reference the video from Data Tutorials in terms of the scope of the project , however, the way my thought process or even some SQL queries will be different : <https://www.youtube.com/watch?v=IaA9YNlg5hM&list=PLNr6y7fJuf_f9wCIPQTun4pMosf5e4fFk&index=2>
2. **THE QUERIES (the one I use and the one Darshil use) WILL BE IN A NOTEPAD/SQL FILE !!!!, THIS IS JUST TO SEE THE RESULT AND QUERY BY PICTURE**
3. This is for learning purposes to get better in Azure

**Theory for this project:**

Pipeline for this project: (ETL: extract transform load)

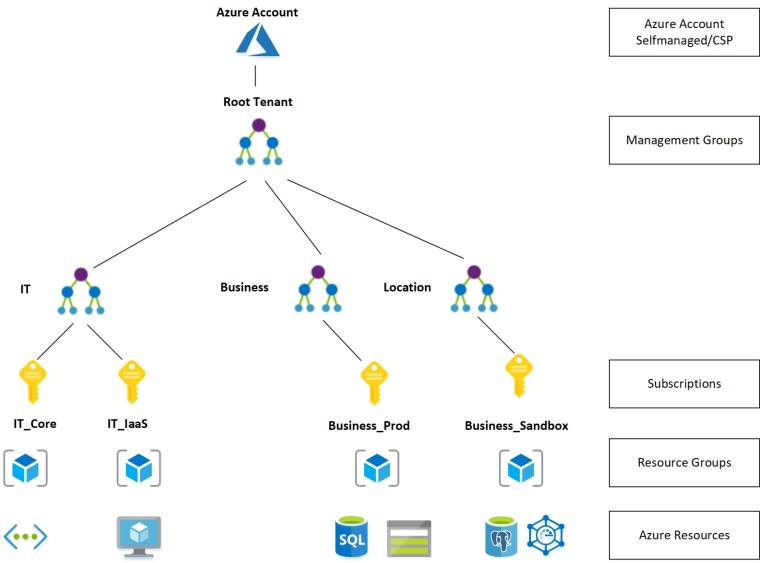
Data Sources -> Data Indigestion (data factory) -> Raw data store -> Transformation (Azure data bricks) -> Transformed data -> Analytics (Azure Synapse Analytics) -> Dashboard (Power BI \*may skip this step)

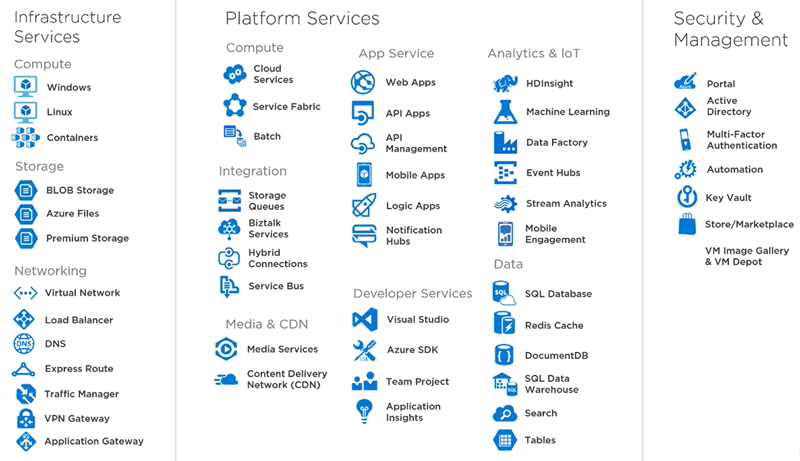
Understand Dataset and API:

1. Contains 11000 athletes, 47 disciplines, 743 teams and its members partaking in 2021(2020) Tokyo Olympics and many other information like gender, country, etc.
2. Multiple different files (Kaggle is in xlsx but we need it in CSV) so just transform the file type.

Different Azure Services and Account:

This will depend on your company or project requirement.





What we will be using for this project:

1. Data Factory
2. Data Lake Gen 2
3. Azure Databricks
4. Synapse Analytics

**SET UP PROCESS**

Step 1: Make an account for Microsoft Azure

Step 2: Store The data in the data lake

* In the search bar type: Storage Account
* Create storage account
* Subscription: Free trail, resource group: create new and rename “Tokyo Olympic”
* Storage account name needs to be unique, region: South East Asia
* Performance as standard, redundancy as GRS (though irl when work will be diff)
* Enable Hierarchical Namespace!
* Keep clicking next till the end then wait till deployment is complete

(“go resources” for more details)

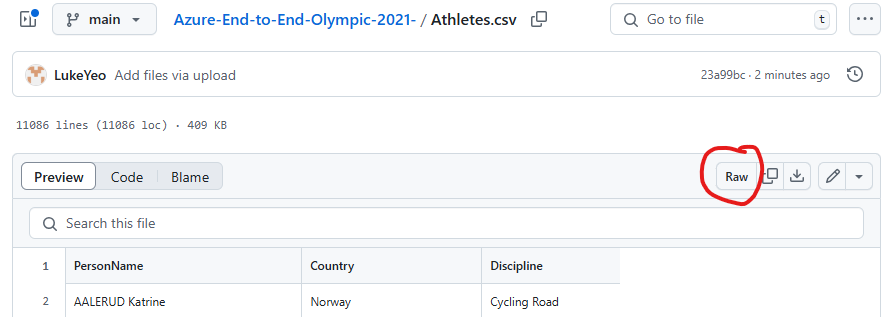
Step 3: Creating folders for the different versions of data

* Go to container under Data Storage
* Add container and name it appropriately
* Inside the container, we want 2 folders so create (add directory): “raw-data” and “transformed-data”

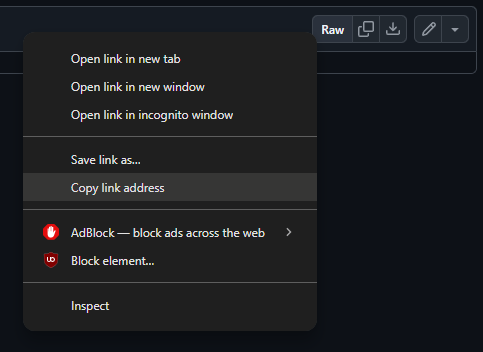
Step 4: Using Azure Data Factory

* Do the same input as the Storage Account
* Name can put Tokyo Olympic Data Factory
* Click “Launch Studio”
* Create a new pipeline and name is as data-ingestion
* “Activities” tab is the main one we will be using, a lot of the services are “linked” together
* In “move and transform”, just drag and drop “copy data” -> source -> sink in the bottom tab
* Source:

1. click the “+” then get the data from “HTTP”, copy paste it from the Github data file itself link (1 data file at a time). So go to each data file and click on “Raw”



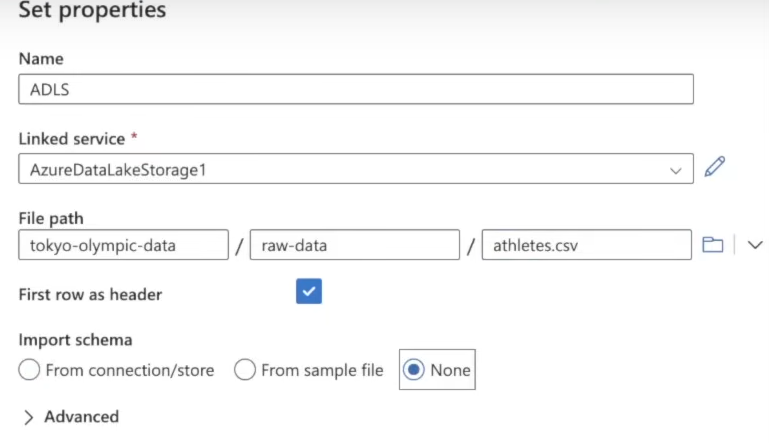
* if this does not work, use the copy link address instead



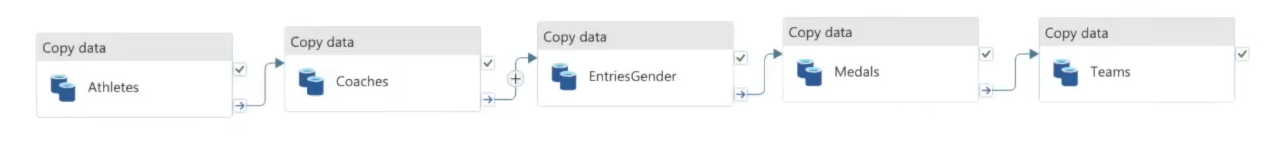
1. Press “linked service”, in base URL, copy paste the whole thing. Authentication type, put anonymous.
2. Tick the first-row header then presses OK

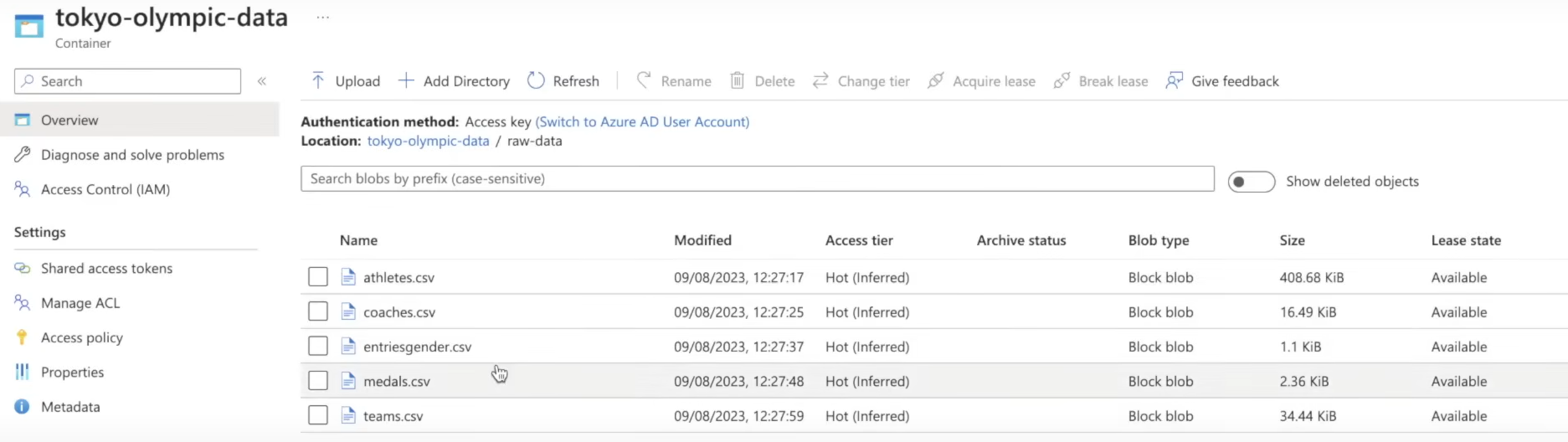
* Sink:

1. When you click the scroll for sink, the datasets should be there already but ignore that.
2. Click on the “+”, then find “Azure Data Lake Storage Gen2” then click CSV file as normal.
3. For linked service, change the azure subscription and storage account name as what we did for Step1 (“Tokyo Olympic”)
4. In file path, click browse and put in into “Raw-data” and put in the file name like “Athletes”
5. Click on the First row as header
6. Import schema: “None”



* Click on “Validate” at the top right after finishing Source and Sink.
* If want to check, press debug button and see if there are any errors then wait for “succeeded”
* Check the original Raw-data folder in step3. **Then do for each data file 1 at a time.**
* Use the arrows in the logo to connect the datasets (debug then wait for success again)



Checking on the “Raw” folder itself

Step 5: Azure Databricks (Data Transformation):

* Same info as Step1, but workspace name can be “Tokyo Olympic data bricks”
* Click “Launch workspace”
* It will seem like a Jupyter notebook workspace
* Go to compute and press “create compute”

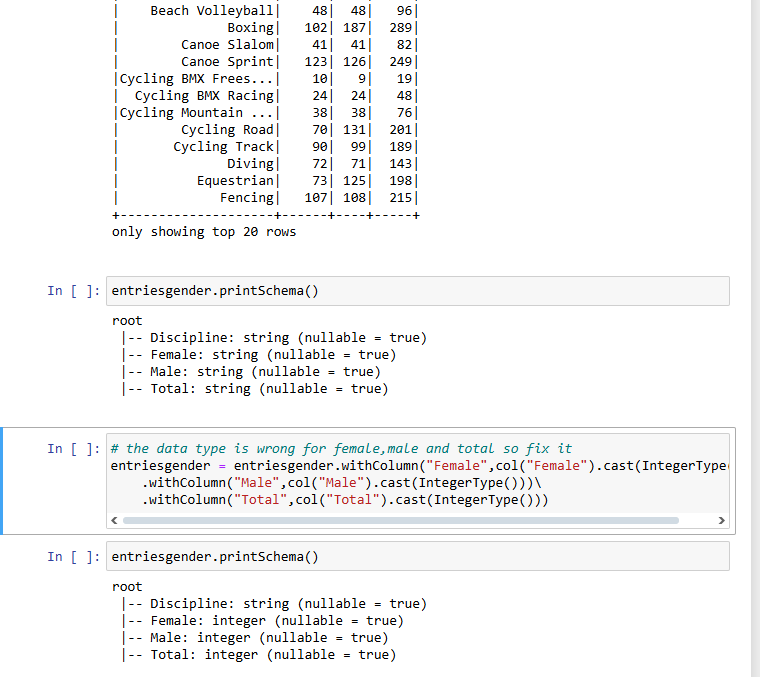
1. Policy: unrestricted since I am working on this alone
2. Put as single node
3. Don’t have to change anything else

* Click on notebook (look at the ipynb file)
* Search for App registrations, name it and create the app.
  + Name: App01
  + Client ID
  + ObjectID
  + DirectoryID
* Client secret: make 1, copy the value and SecretID

Note: instead of saving it in a notepad, azure has a key vault that we can use to save important ID and info.

* For access, go to the data storage folder, click on the data and then click Access Control (IAM) , add “role assignment” then click on “storage blob data contributor”.
* Click select members, there should be an option for name, put the correct name (App01) then click next till the end.

Example of the transformation:



Standalone Step: Synapse workspace (All in 1 Services)

* Create the synapse workspace, same as step 1
* There is where we can do all the things we did from stop 1-5

Step 1: Click on late data, name it as “TokyoOlympicDB”

Step 2: Click on datable then “From Data Lake”, Tick the infer column names

Step 3: Then click validate and publish

Step 4: Need to refresh each time you add in a new dataset

* On the data -> table, check the “use database on the top tab” if it is correct.
* Can right click to use SQL to do any analysis
* Once we put in the SQL, there is a Table or Chart tab that we can switch between for visualization