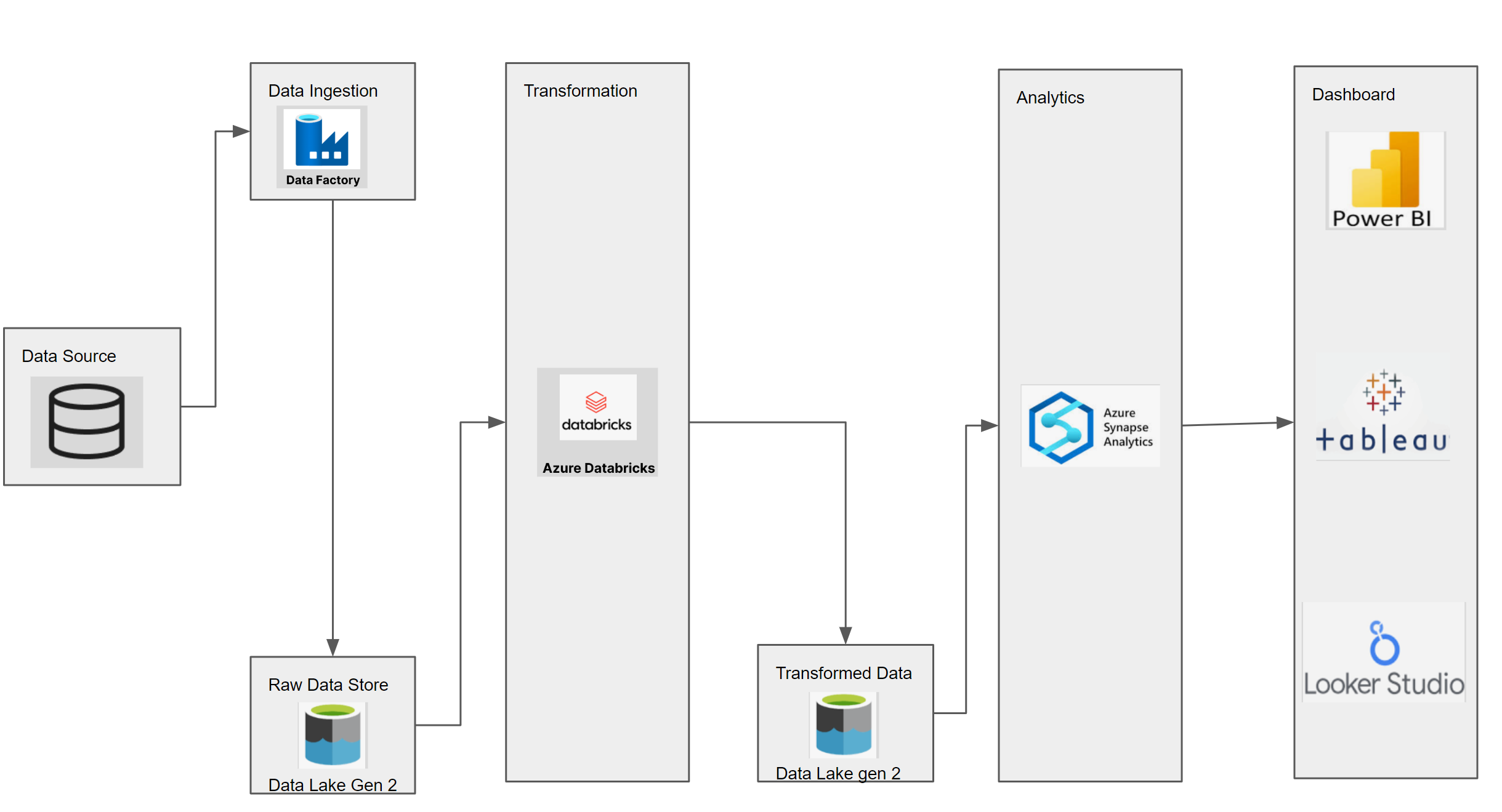
**Flow Diagram:**



**Steps to create storage account & container:**

1. Create Storage Account:

* Give Resource group name.
* Give storage account name. Select region.
* Performance: Standard for now
* Enable hierarchical namespace.

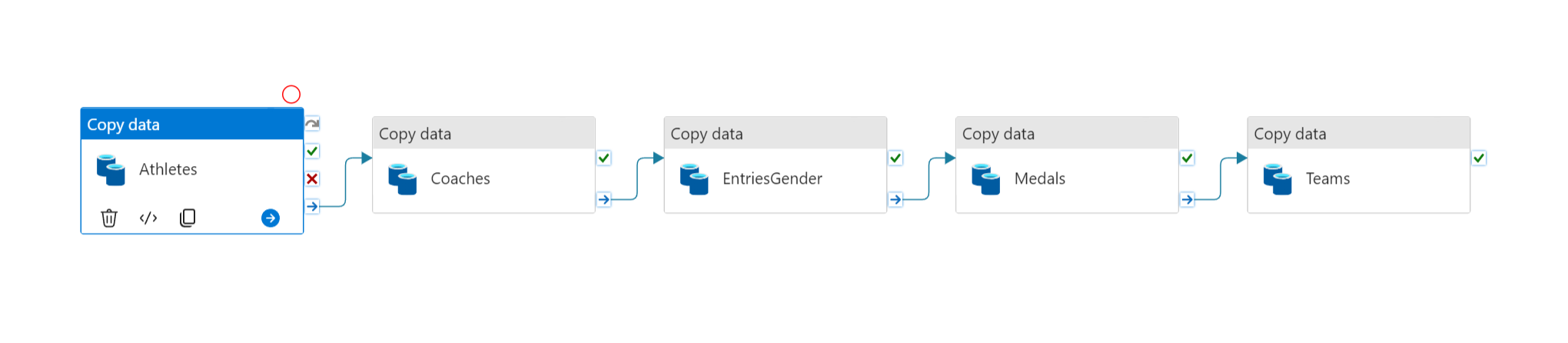
1. Click on go to resources after storage account is created.
2. Click on Containers and add a new container.
3. Add directory:

* Raw Data
* Transformed Data

**Steps to create Data factory & Data flow:**

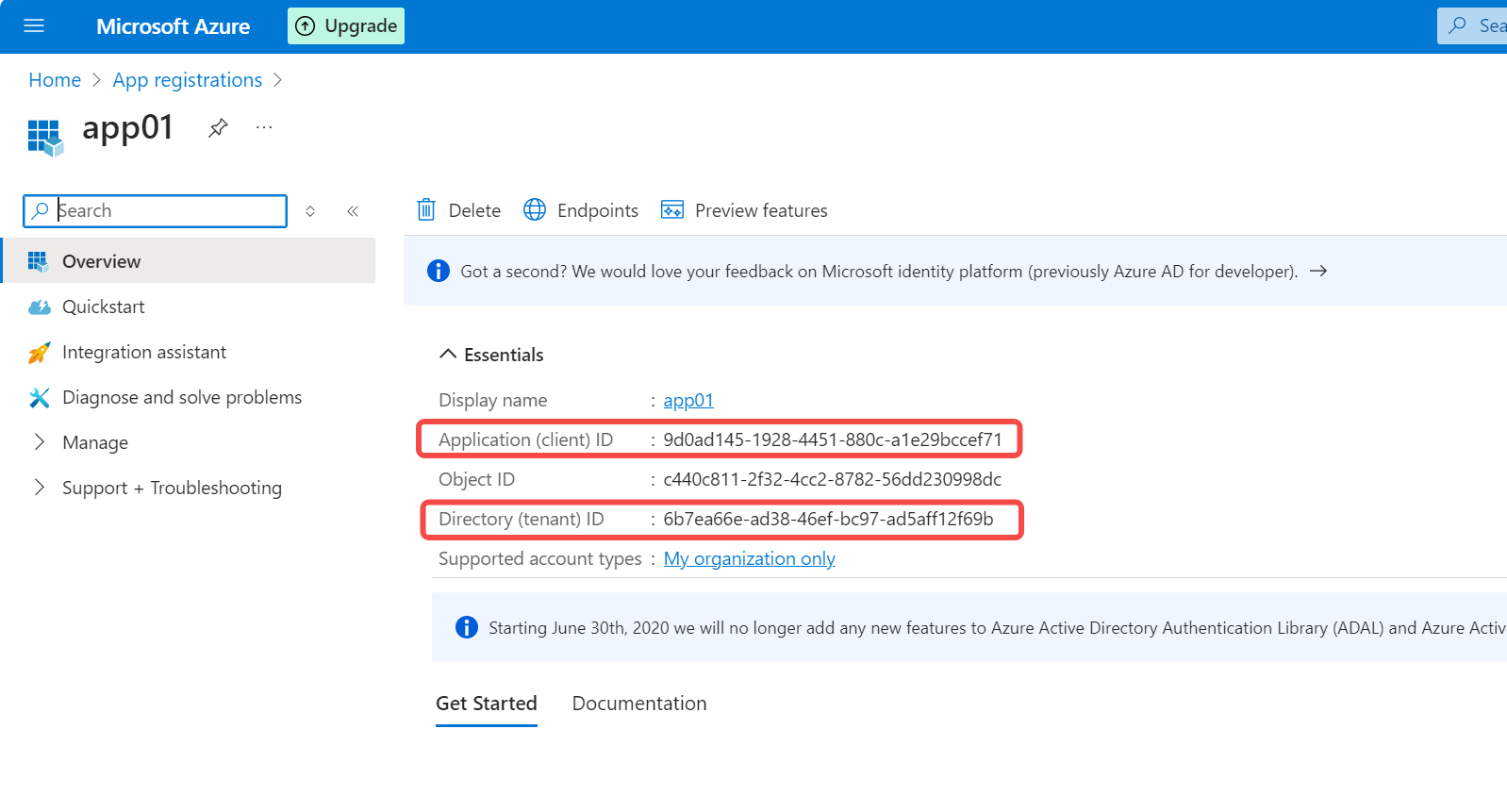
1. Search for data factory, create data factory.
2. Select resource group.
3. Give name, select region.
4. Review and create.
5. Go to resources. Launch Studio
6. Go to Author and create pipeline 🡪give name.
7. Under activities click on move and transform and select Copy Data. (Keep your raw data files on GitHub repository, its easy to link GitHub repo. For data copying) otherwise you need to get access to API & authentications keys etc.
8. Go to GitHub repo having files and copy raw URL.
9. Under Copy data 🡪Source section click on + icon and search for HTTP and select it
10. Select CSV file format and then fill in new linked services info and paste the above copied raw URL from GitHub repo.
11. Click on Preview Data to see the data.
12. Now we need to configure Sink 🡪 Create new source 🡪 select Gen2
13. Select CSV format.
14. Set properties 🡪 create new linked service.
15. Set file path 🡪select raw data directory🡪Import schema will be none.
16. Debug the copy data task and see if the data is copied under raw data directory.

**Repeat above activity to copy all the other different files present in our GitHub repository.**

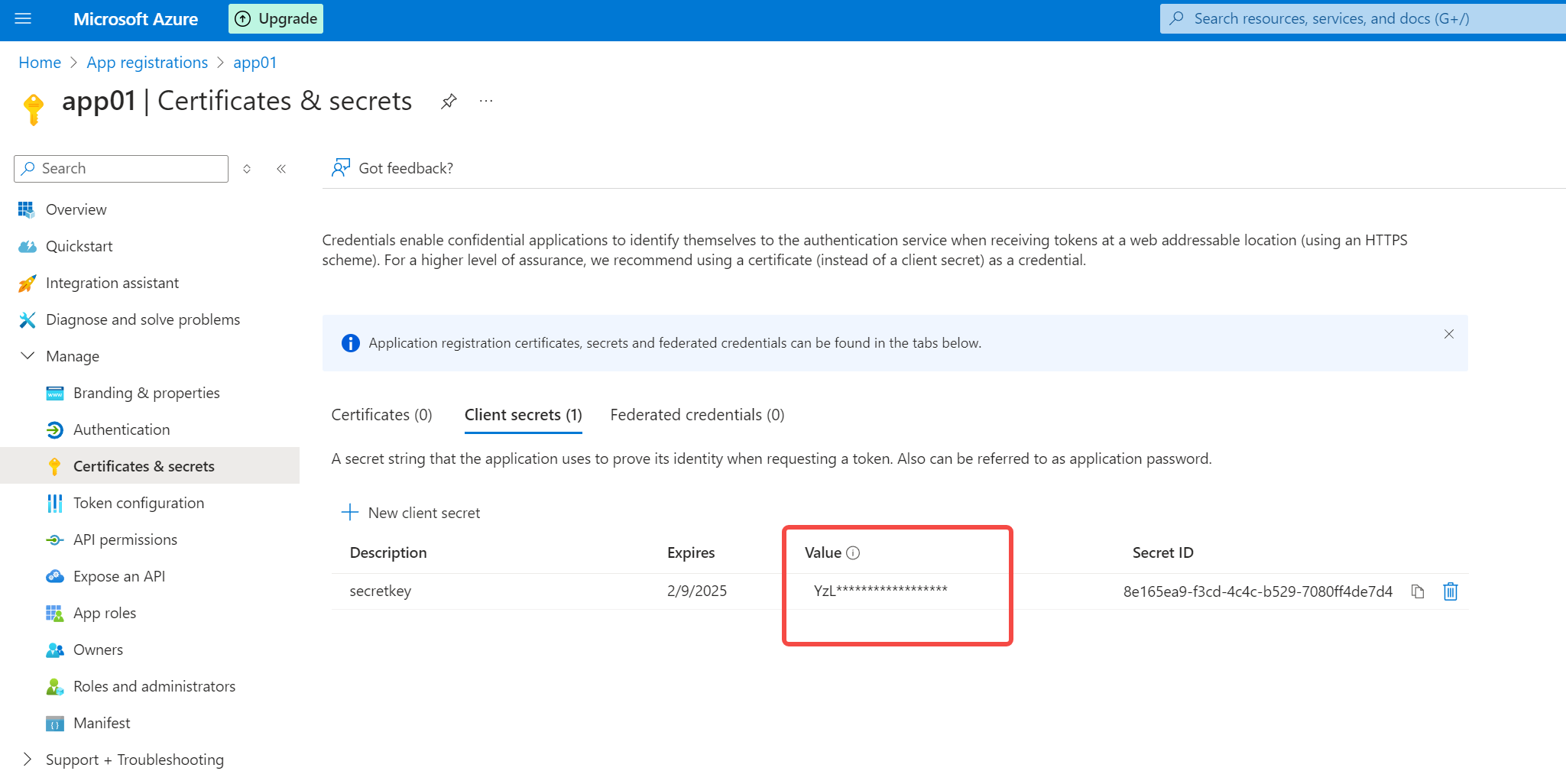


**Steps to perform Data transformation:**

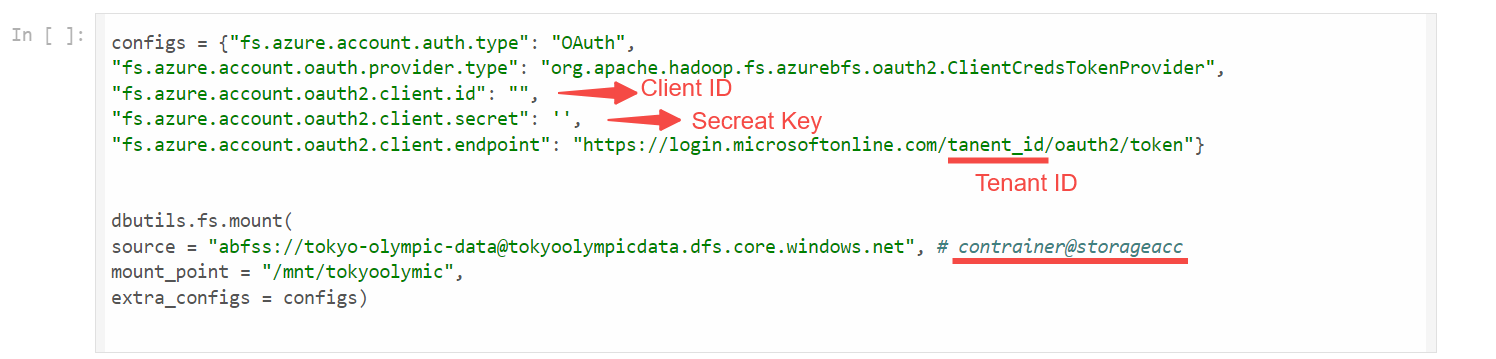
1. Search for Databricks🡪 Create an azure databricks workspace.
2. Create Compute, (spark code runs on something)🡪 single node & user🡪 create compute.
3. Go to resource🡪 launch workspace🡪New Workbook (Tokyo Olympic transformation)
4. Mount the data lake storage with data factory.
5. Search for App registration🡪click on it and give name 🡪 Copy Client ID and Tenant ID



1. Click on certificates & secret🡪New Client secret🡪give name🡪 copy value.



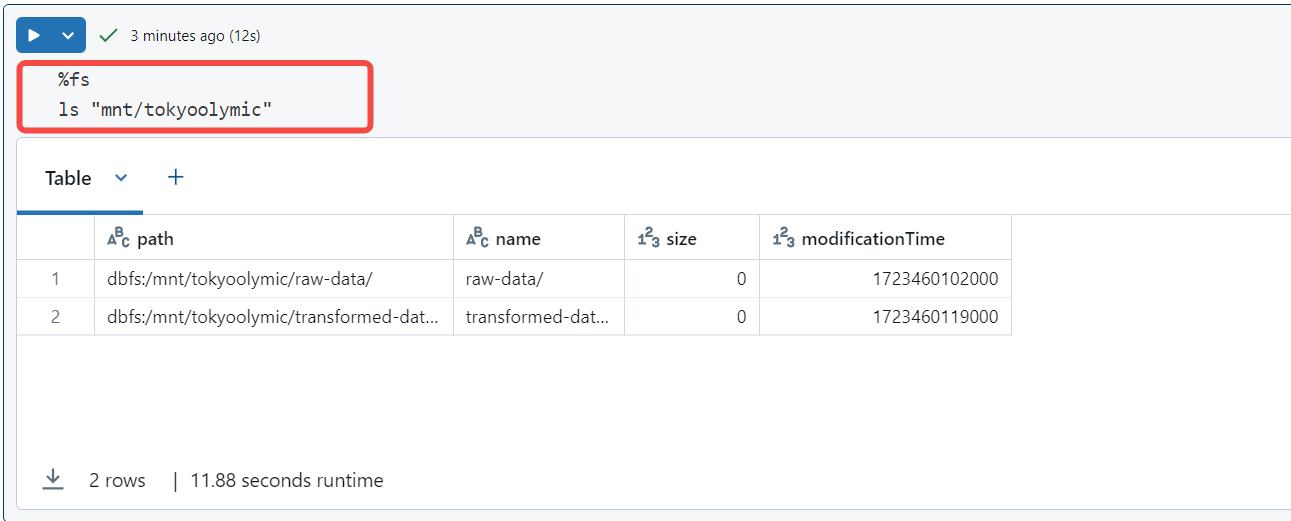
1. Start writing the code in the notebook:



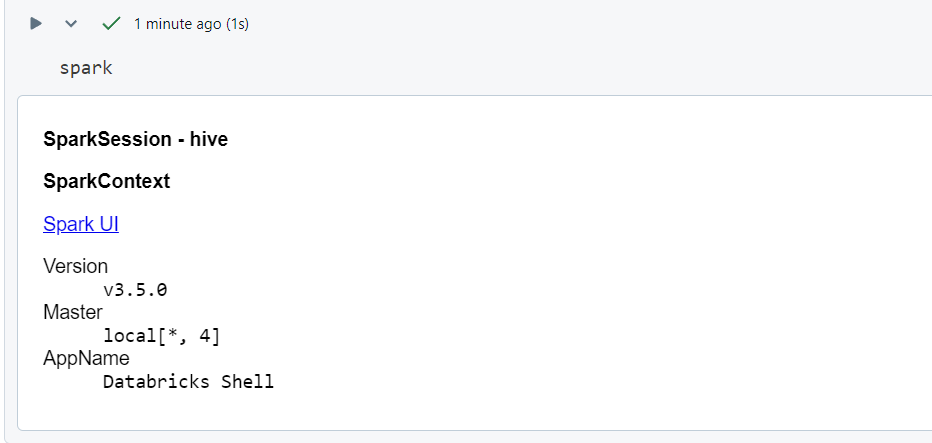
**Sometimes you get error: request is not authorized to perform the operation due to insufficient permissions.**

**Go to Storage account🡪click on container 🡪Click on Access Control (IAM)🡪Add role assignment 🡪Give access to “Storage Blob Data Contributor”🡪Click Next 🡪Select member’s search for your app and add it.**

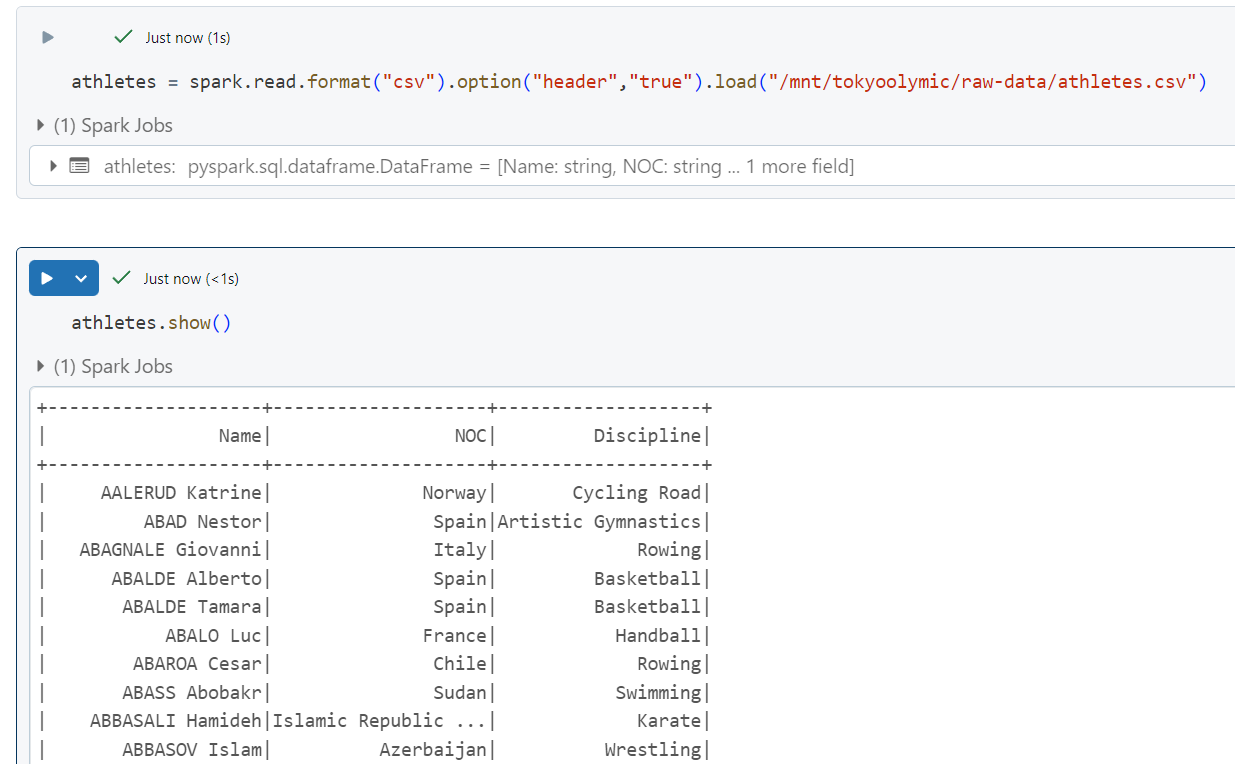
1. Run below code to check if the mounting is done.



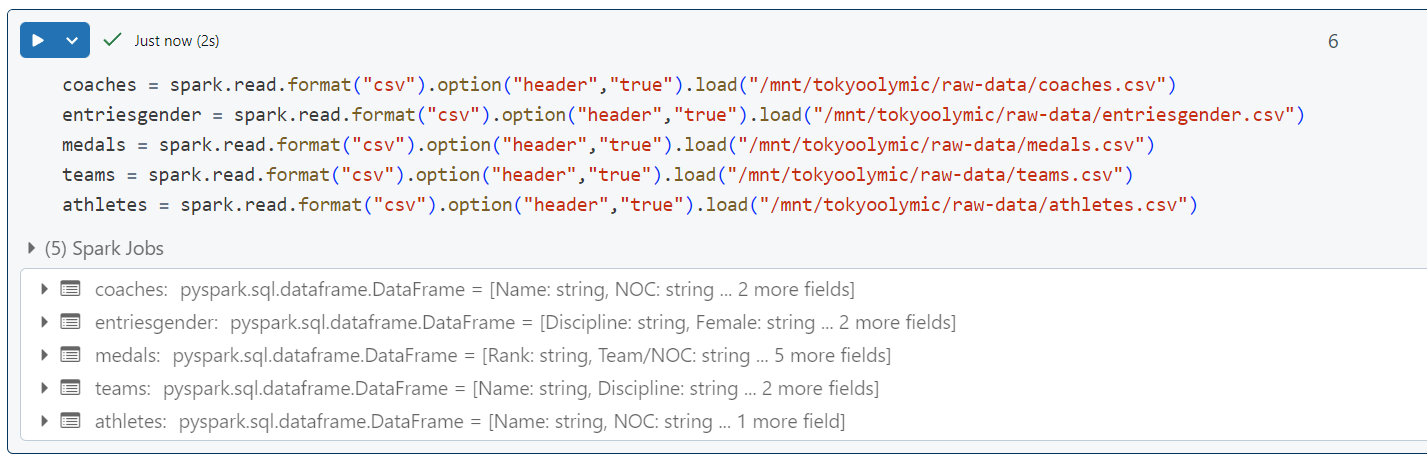
**Note: Generally when you write spark code you need to generate Spark Session , create the app in spark session and then write code in it , but in Databrick you already get spark session created by default you don’t have to explicitly create it .**



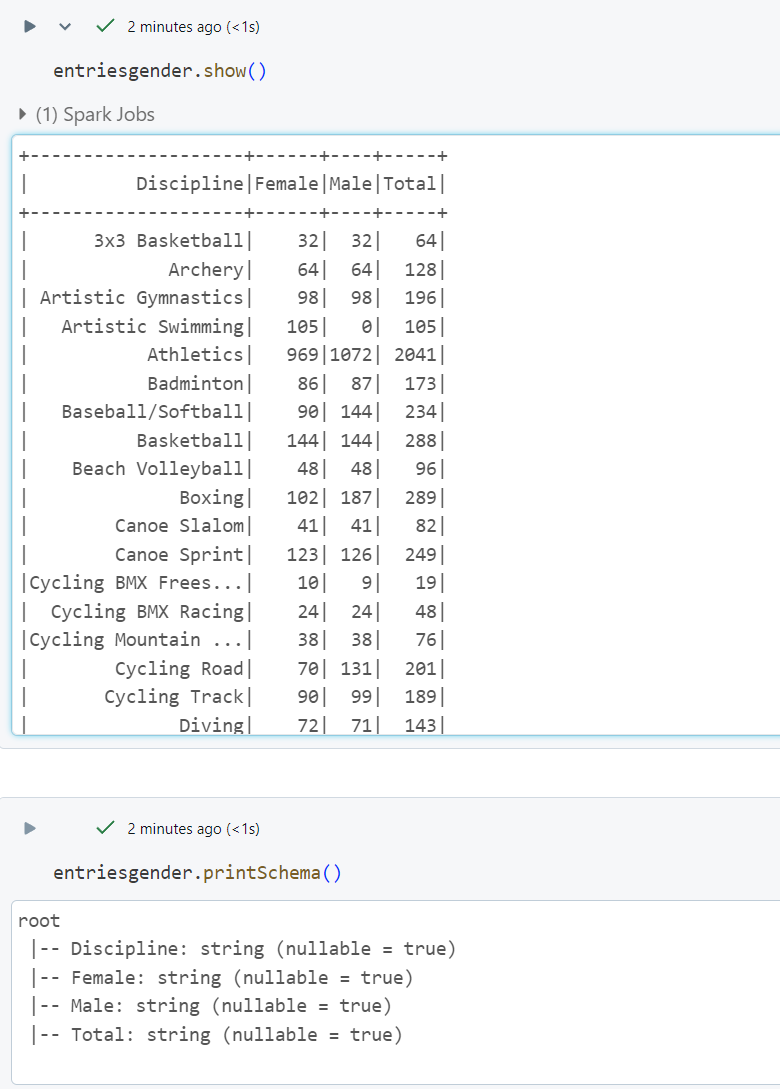
1. Now we will try to read file data:



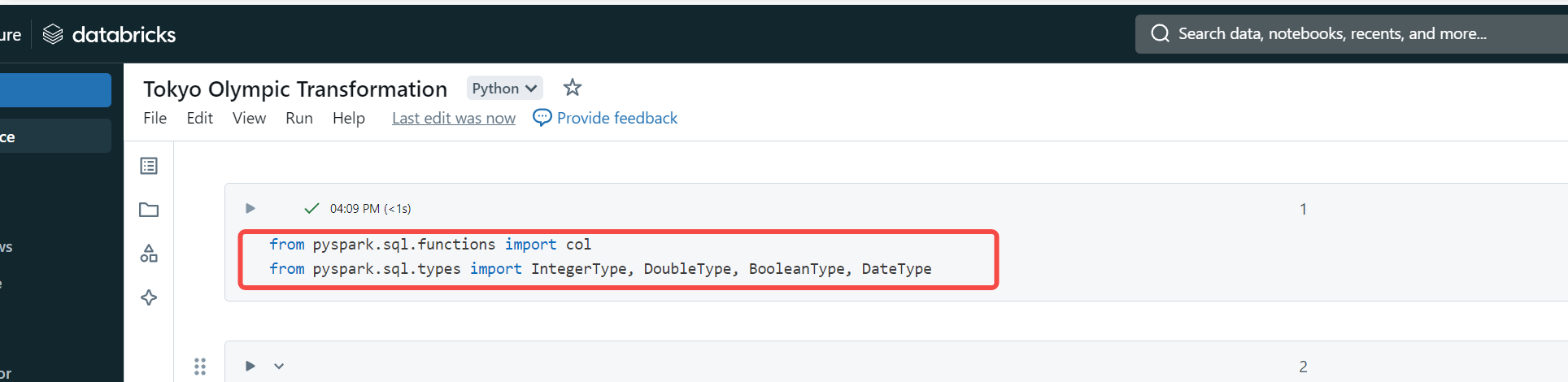
1. Now we will try to read all the files from our raw data folder in the container:



1. Now if we look at the entriesgender table and do a show () and look at the schema, then we can see the Female, Male and Total column data type is showing string although it contains integer values.



1. Now we need to change data type of these columns, so to do that we need to import column function and Integer type function in spark first:

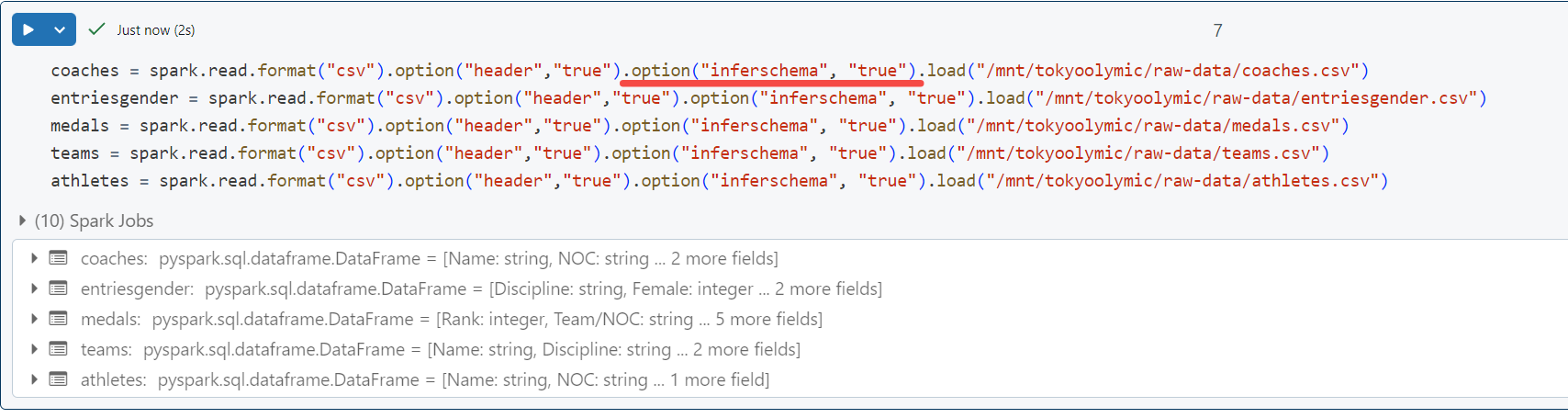


The write the code to convert the data type.



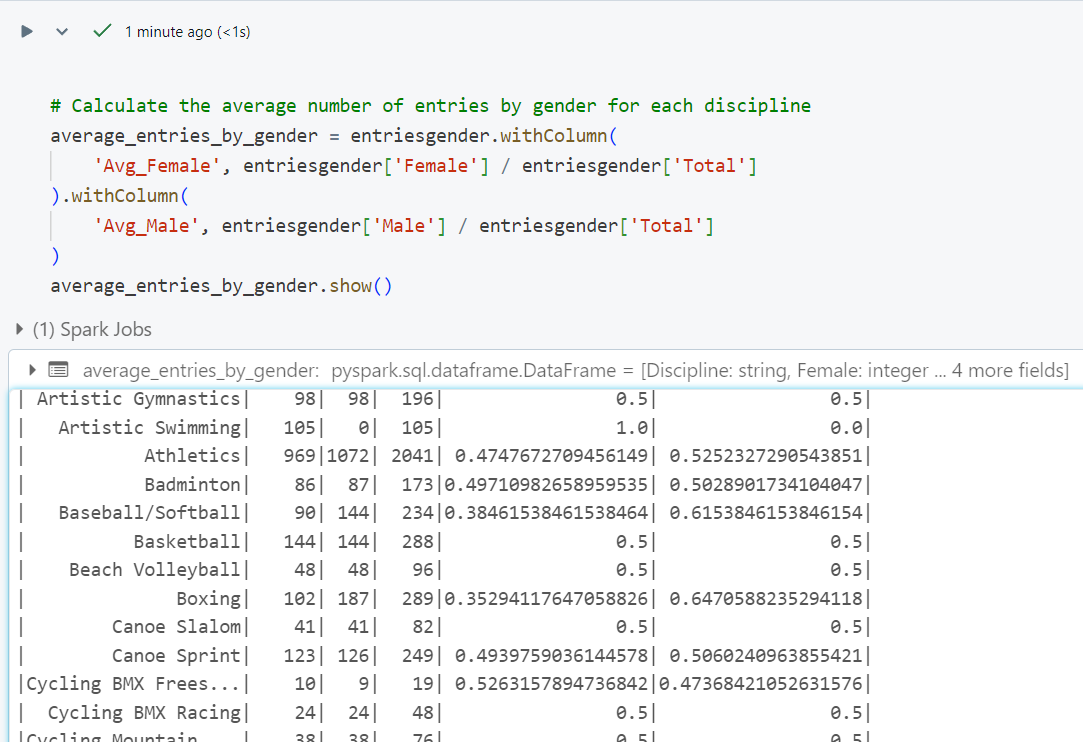
Now we do the same for medals table .

We can do it manually by writing the code like above or else we do have other way where we use **.option(“inferschema”,”true”)**

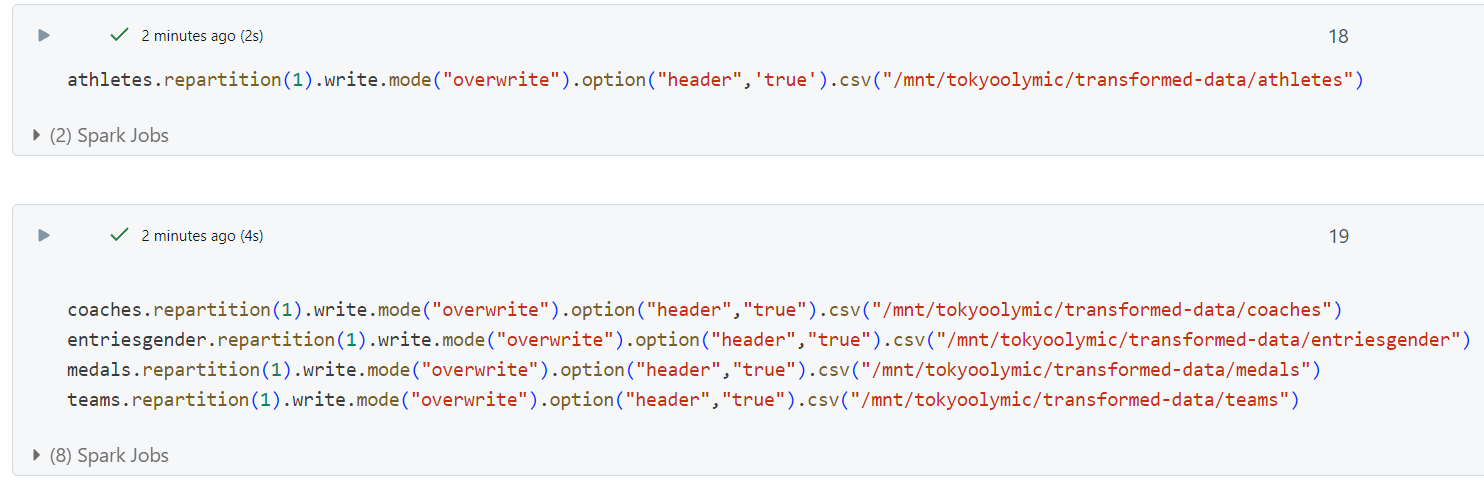


1. Now let’s perform some transformation on the data:



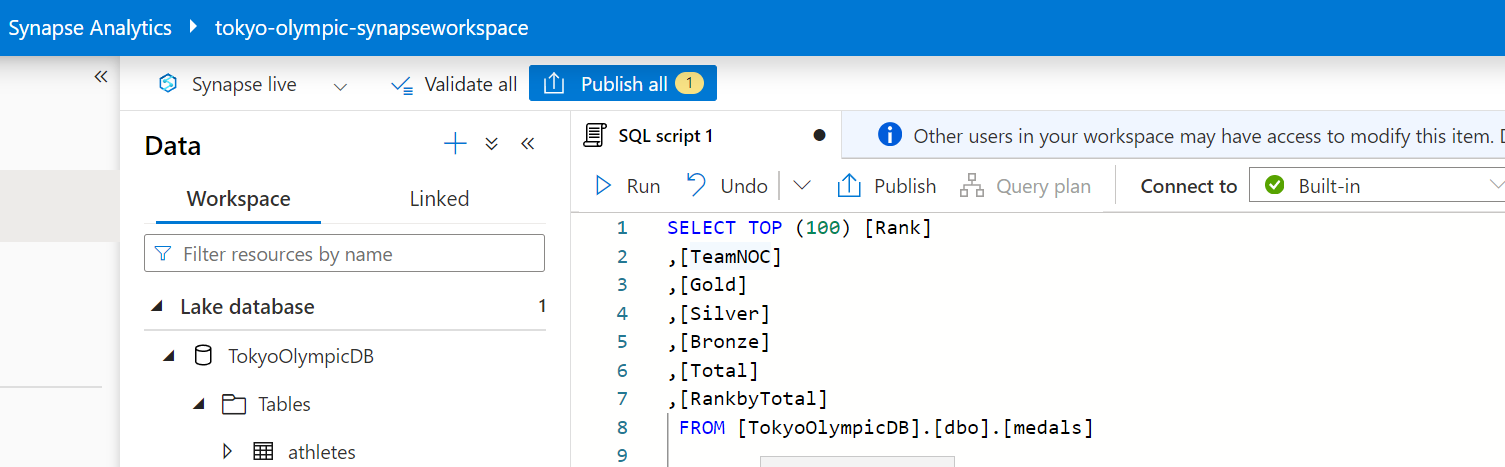


1. Now we must write the data to transformed data folder:



**Performing Analytics using Azure Synapse Analytics**

1. Search for Synapse Analytics🡪Create workspace🡪Account select Olympic data🡪review and create
2. Click on got to resource🡪Click on your synapse workspace which you created above
3. Open Synapse Studio
4. Click on Data 🡪 Click on + icon and select Lake database🡪give DB name
5. Click on +Table 🡪give name 🡪linked service select default 🡪select the file from the raw data folder
6. Select infer column name
7. Click on validate all🡪then publish 🡪refresh database and you will see the table.
8. Now you can query data from the table .



Now you need to create table for all the raw data files as above.

