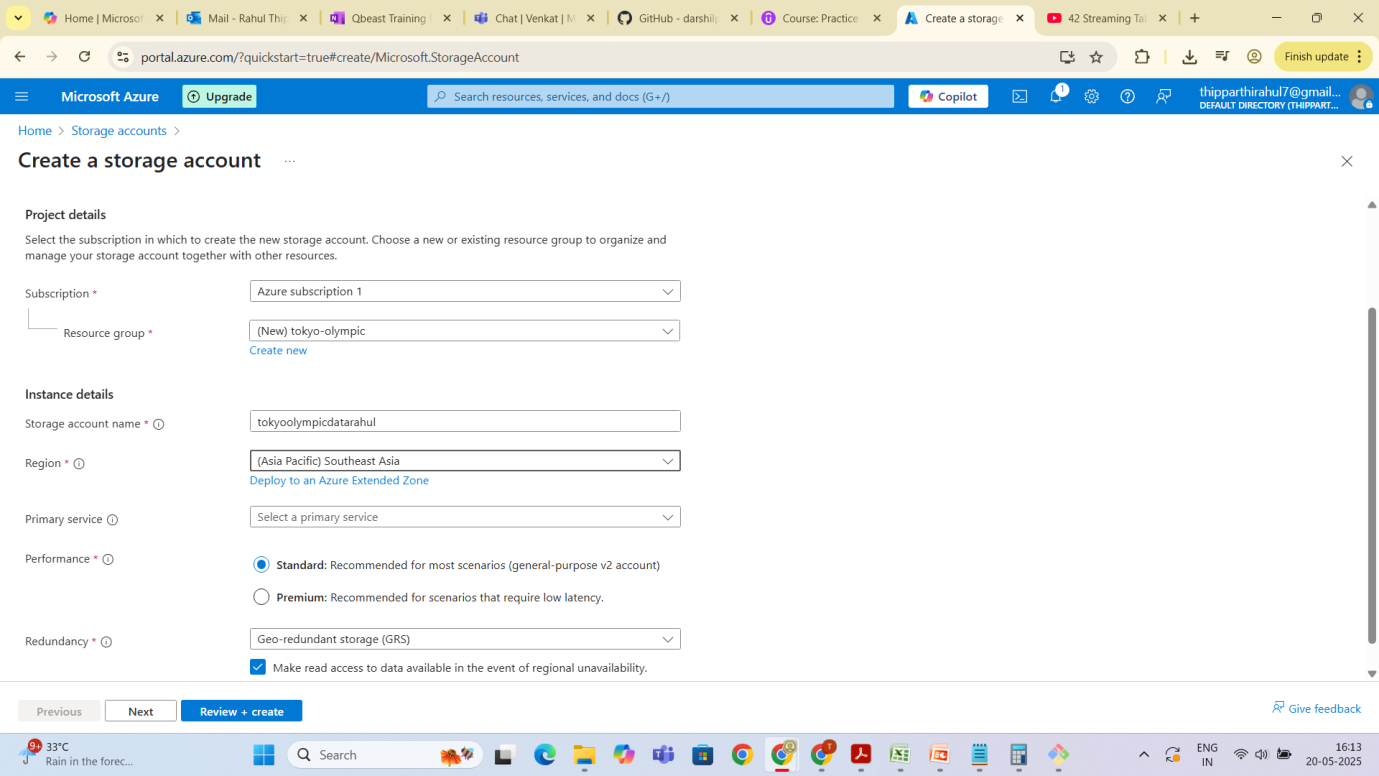
**📘 Azure Data Engineering Project: Tokyo Olympics Data Pipeline**

**Objective:**

To build an end-to-end data pipeline using Azure services that ingests raw data from GitHub, processes and transforms it using Azure Databricks, and stores it in Azure Data Lake and Azure Synapse Analytics for analytics and reporting.

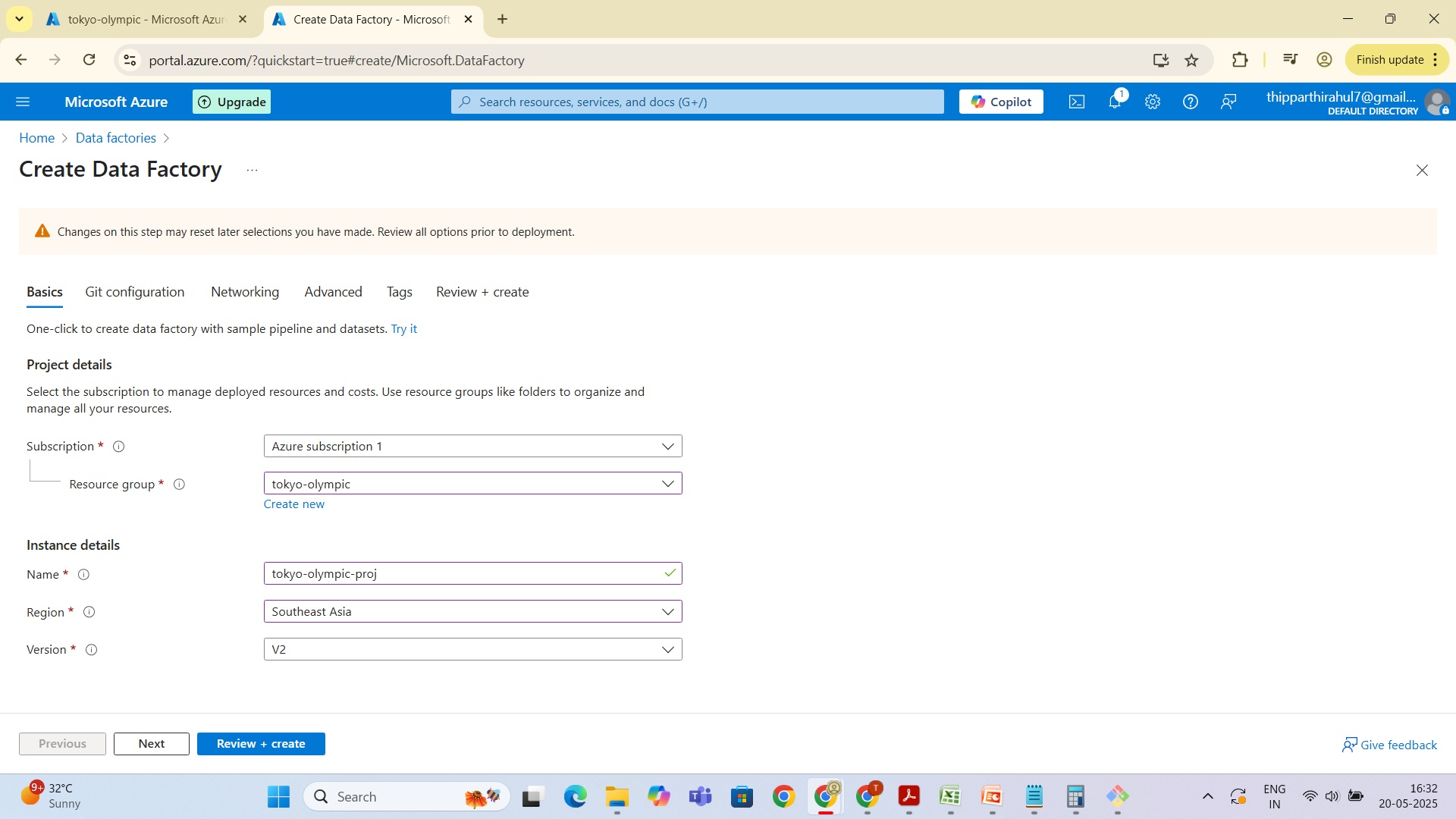
**🔷 Step 1: Azure Data Lake Storage Setup**

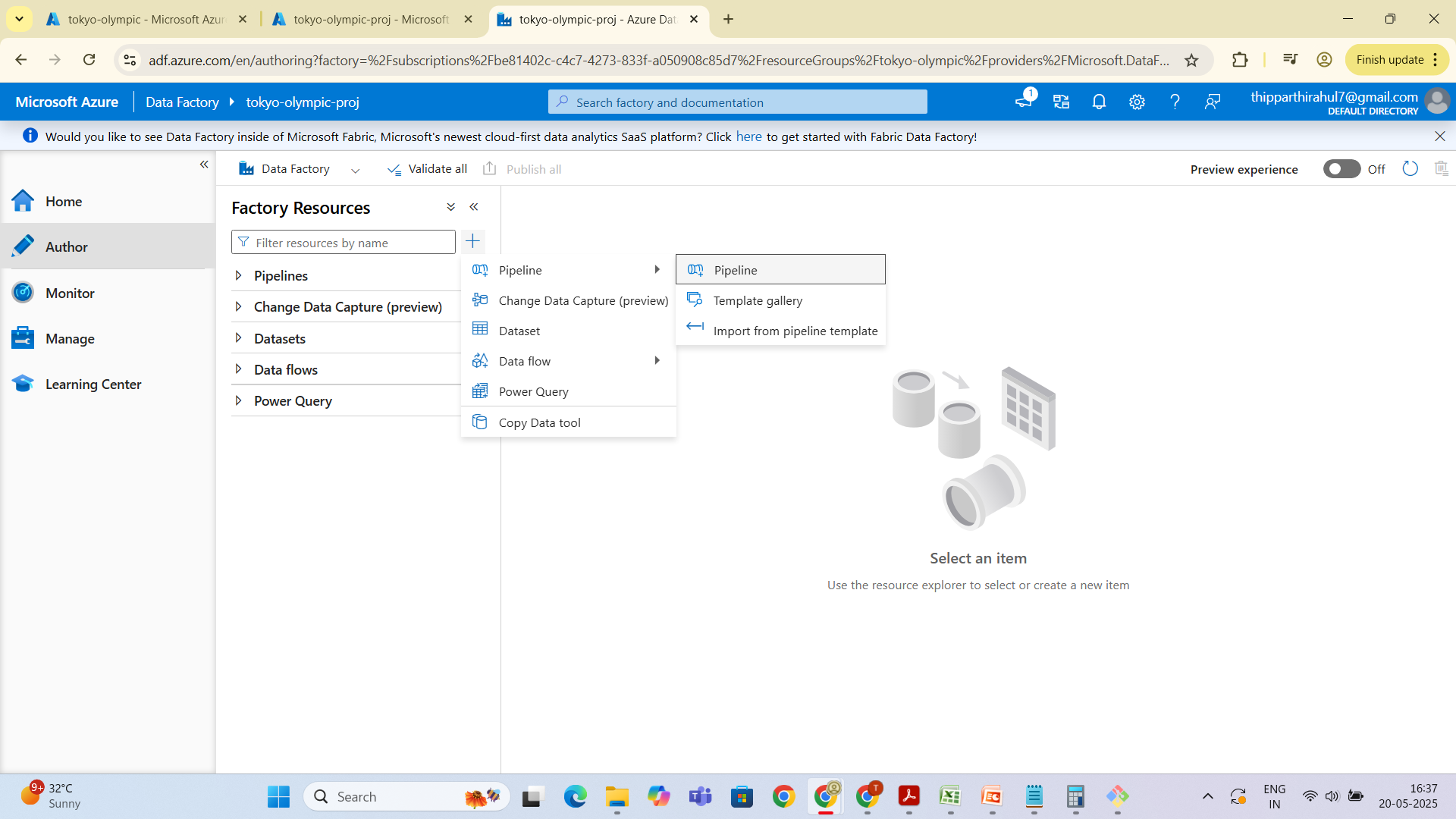
1. **Created a Storage Account** in the Azure portal.
2. **Enabled Hierarchical Namespace** for ADLS Gen2 capabilities.
3. **Created Containers**:
   * raw-data – to store ingested raw files.
   * transformed-data – to store processed data.
4. **Created Directories** inside containers for organizing data files.

****

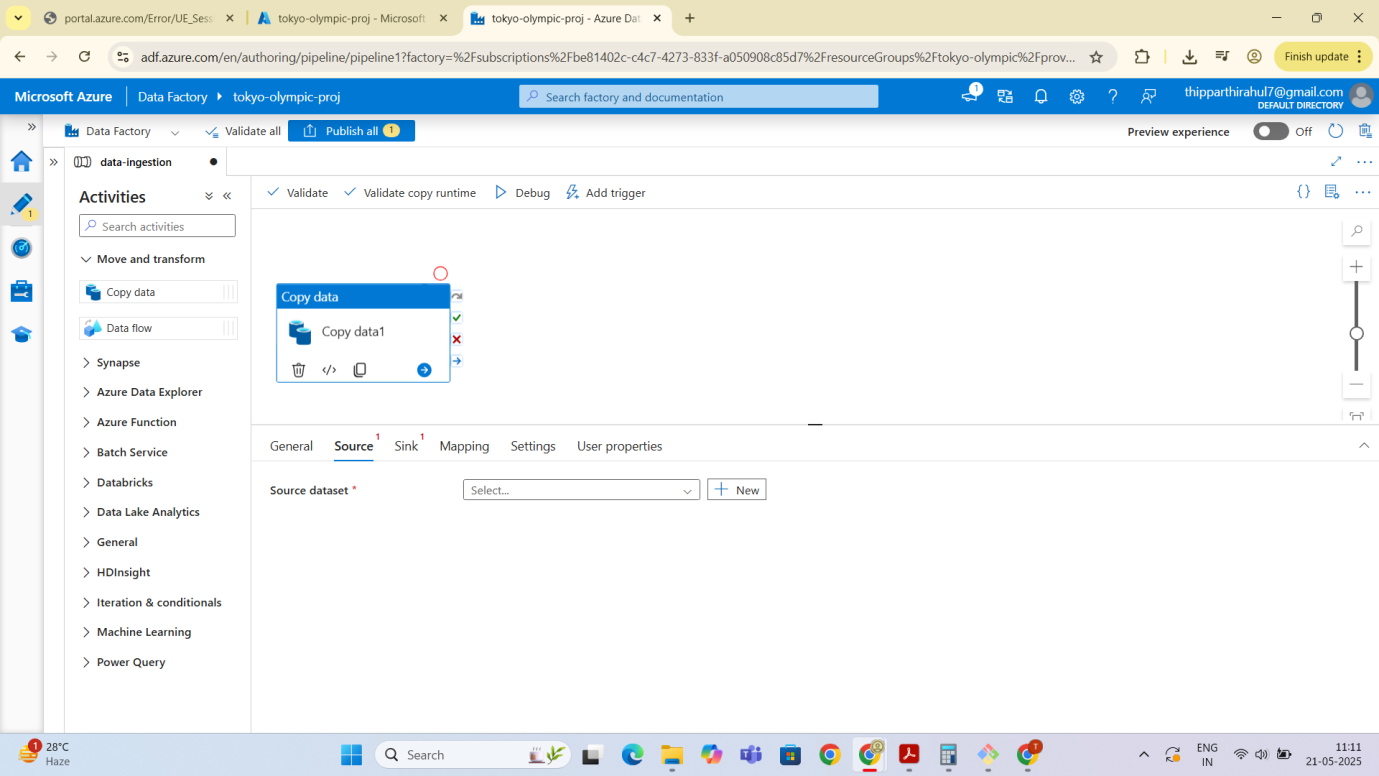
**🔷 Step 2: Azure Data Factory (ADF) – Ingest Raw Data**

1. **Created a Data Factory** instance.

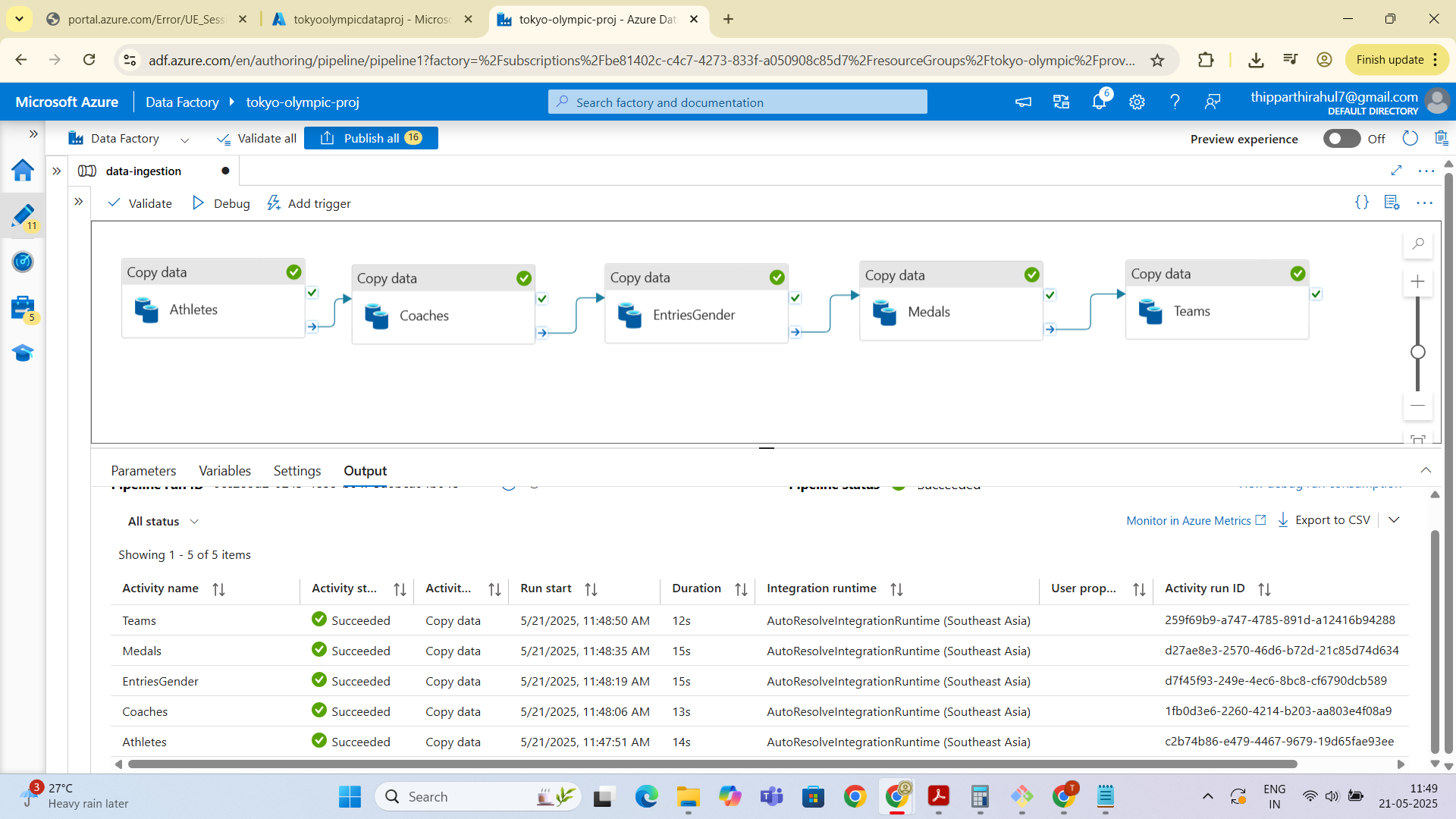


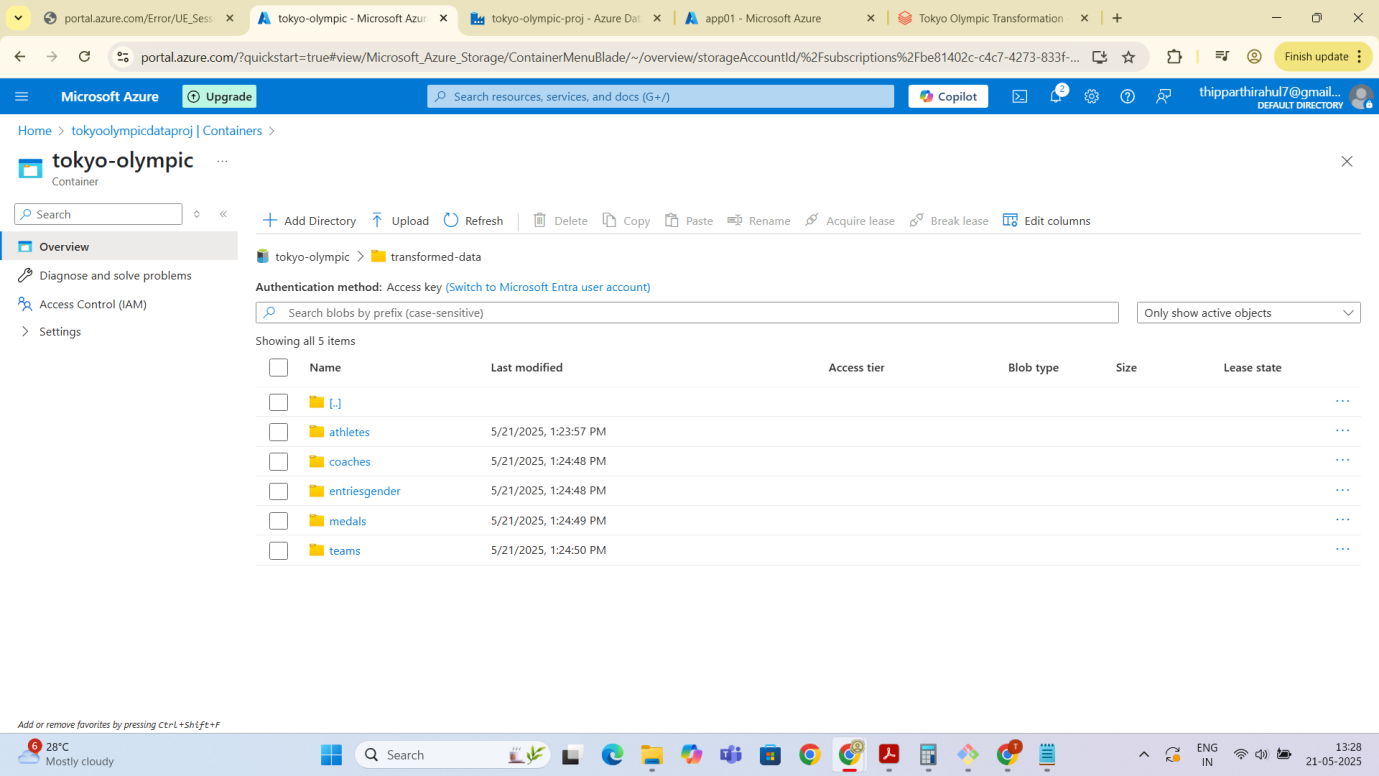


1. **Built a Pipeline** to automate data ingestion:
   * **Added Copy Data activity**.
   * **Source**:
     + Chose HTTP linked service.
     + Provided **GitHub Raw URLs** for datasets.
     + Configured format as CSV or JSON (as per source).
   * **Sink**:
     + Targeted **Azure Data Lake Storage Gen2**.
     + Selected appropriate container and path (e.g., /raw-data/filename.csv).



1. **Repeated the process for multiple files** as required.





**🔷 Step 3: Azure Databricks – Data Processing**

1. **Created an Azure Databricks workspace and cluster**.
2. **Configured the environment for processing** using Python notebooks.

**🔷 Step 4: Azure Active Directory – App Registration**

1. **Registered an Application (App01)** in Azure AD.
2. **Captured credentials**:
   * **Client ID**
   * **Tenant ID**
3. **Generated a Client Secret** under *Certificates & Secrets*.
4. **Copied the Secret Value** for later use.

**🔷 Step 5: Mount Azure Data Lake to Databricks**

1. **Used credentials to configure OAuth authentication**:

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": "<client-id>",

"fs.azure.account.oauth2.client.secret": "<client-secret>",

"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/<tenant-id>/oauth2/token"

}

1. **Mounted the Data Lake into Databricks**:

dbutils.fs.mount(

source = "abfss://tokyo-olympic-data@tokyoolympicdata.dfs.core.windows.net",

mount\_point = "/mnt/tokyoolymic",

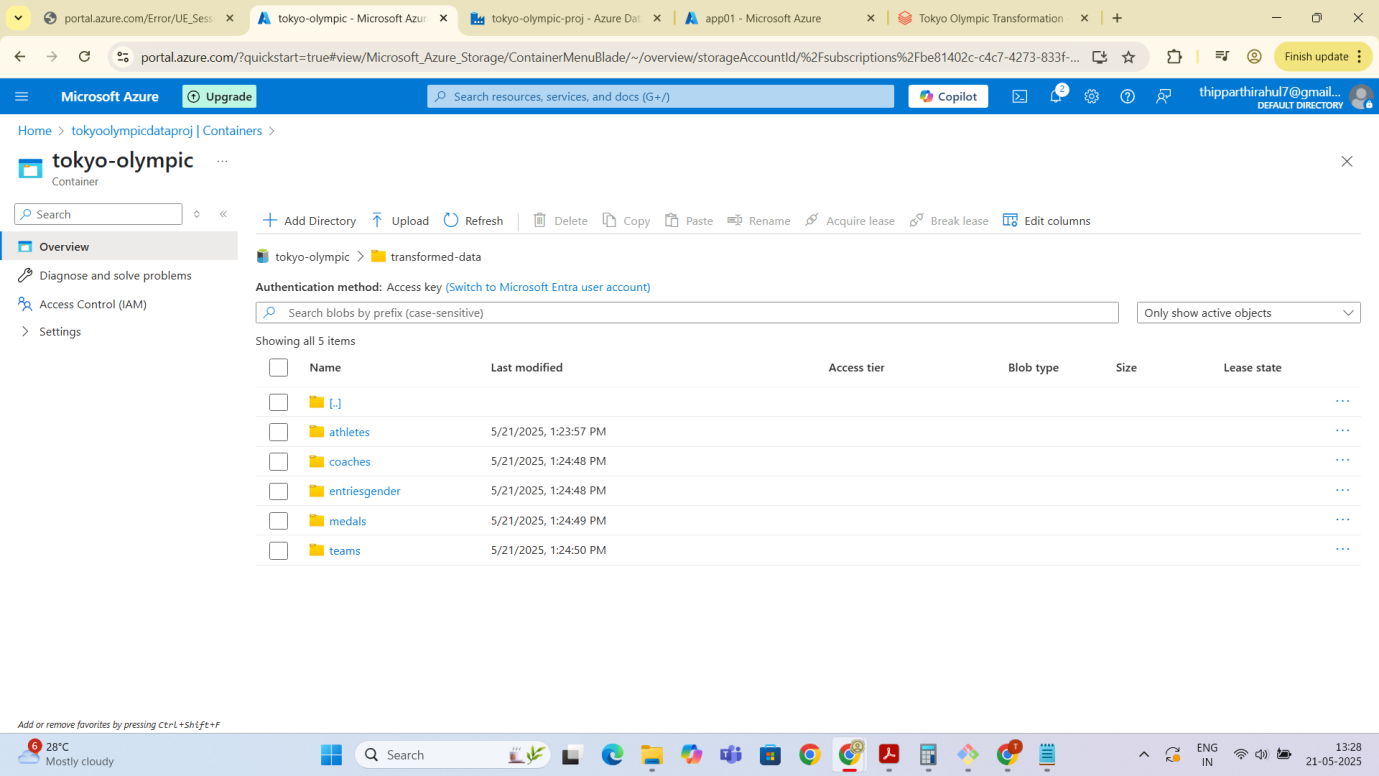
extra\_configs = configs)

**🔷 Step 6: Assign Storage Permissions**

1. **Navigated to Storage Account > Containers > Access Control (IAM)**.
2. **Assigned Required Roles**:
   * Reader
   * Storage Blob Data Contributor
   * Assigned to the registered application (app01)

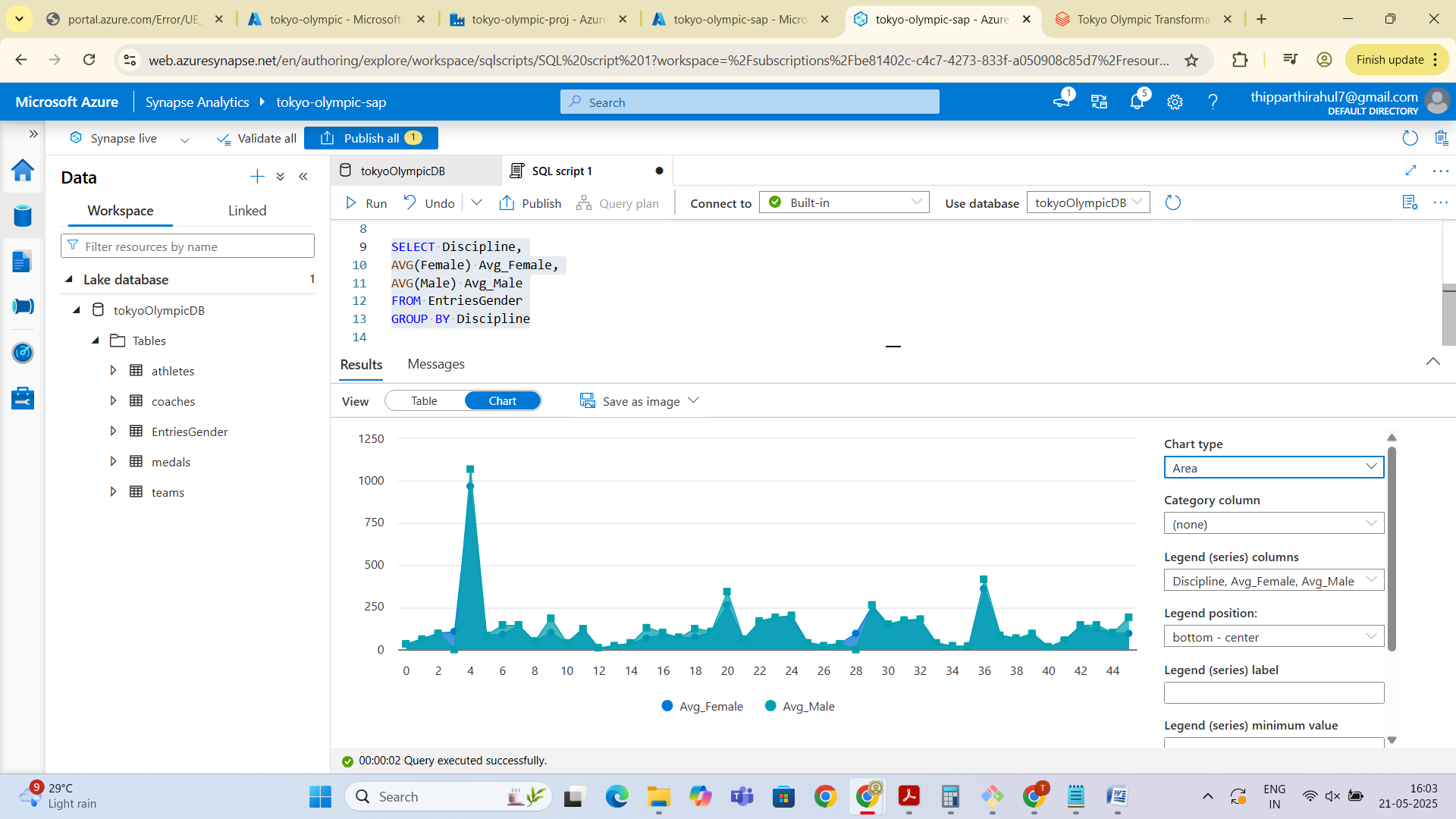
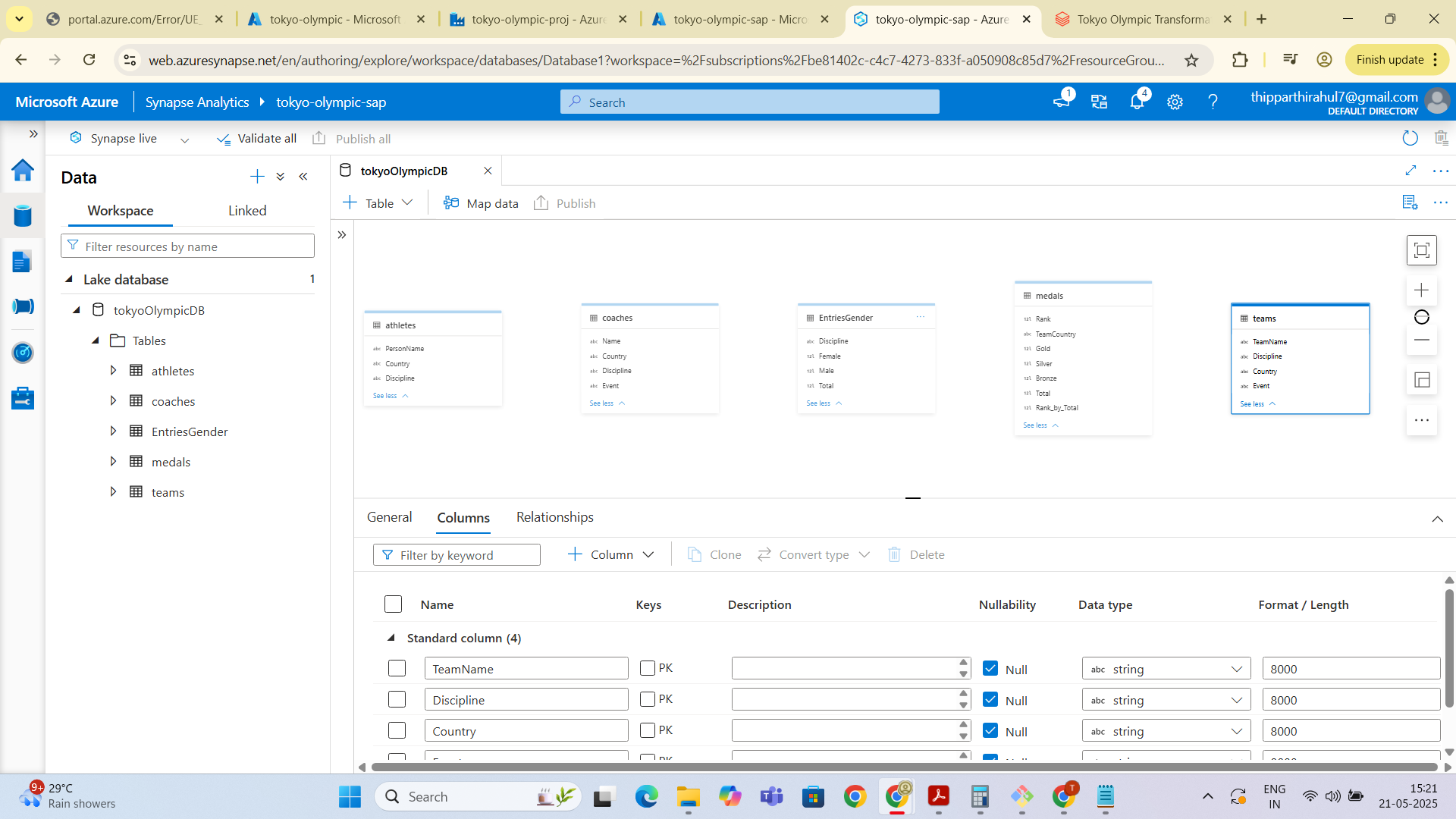
**🔷 Step 7: Data Transformation in Databricks**

1. **Loaded the raw datasets from mounted storage**.
2. **Performed cleaning and transformations** using PySpark.
3. **Saved the processed files to the /transformed-data directory**.



**🔷 Step 8: Azure Synapse Analytics – Final Destination**

1. **Created an Azure Synapse workspace**.
2. **Navigated to Data > Lakehouse > Table > From Data Lake**.
3. **Imported the transformed data files** from ADLS.
4. **Created tables and models for querying and reporting**.



**Outcome:**

An automated and scalable data pipeline was successfully built using the Azure ecosystem, integrating GitHub data, Azure Data Lake Storage, Azure Data Factory, Azure Databricks, and Azure Synapse Analytics.

Let me know if you'd like this exported as a Word or PDF file or added to your portfolio.

