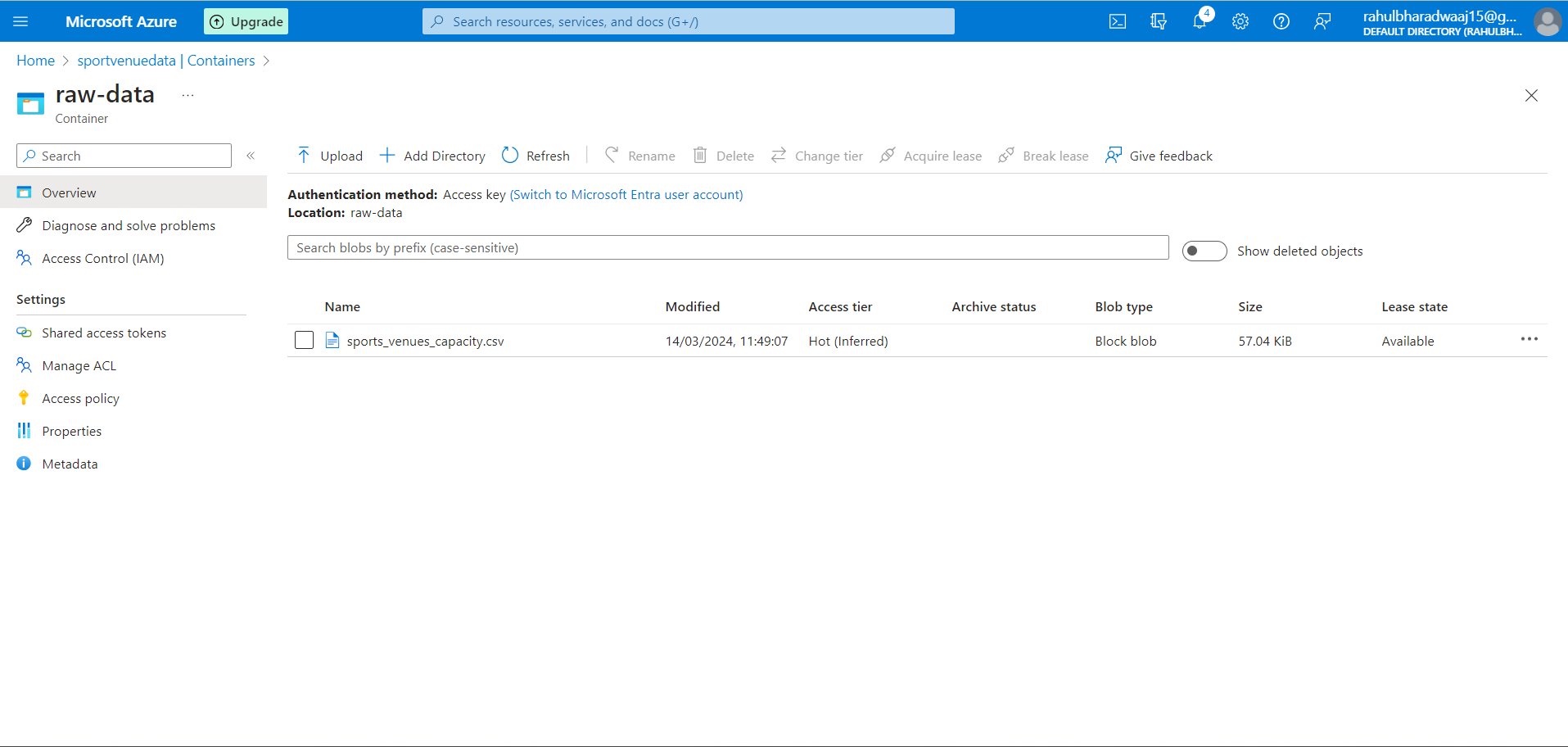
**Sports Venue Data Analysis**

End-to-End Azure Data Engineering Project

**Step 1 : Scrape the data from** [**https://en.wikipedia.org/wiki/List\_of\_sports\_venues\_by\_capacity**](https://en.wikipedia.org/wiki/List_of_sports_venues_by_capacity) **using BeautifulSoup.**

**Step 2 : Store this raw, untouched data to Azure Blob**

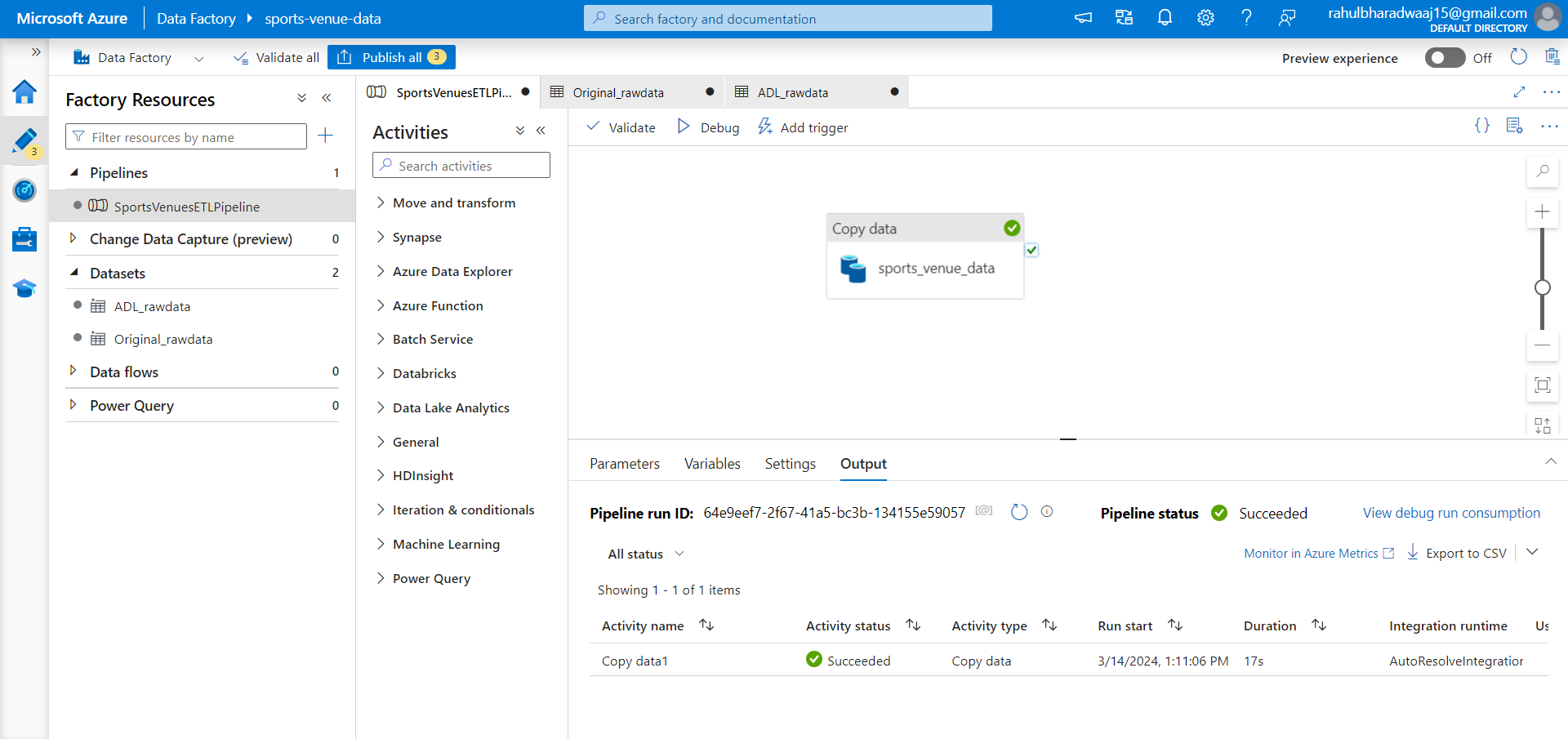
1. Azure Storage account is created with Hierarchical Namespace enabled to leverage ADLS Gen2 features.
2. Used the Python script with BeautifulSoup for web scraping and the Azure Blob Storage SDK to upload the data directly into a container (raw-data) in Azure Storage account.



1. This container will hold the raw data exactly as it's scraped from Wikipedia. Storing raw data separately is a best practice because it ensures we have an unaltered copy of your original data, which can be valuable for re-processing or auditing purposes.

**Step 3 : Transfer Raw Data to a Hierarchical Namespace using Azure Data Factory**

1. Created an Azure Data Factory instance in Azure portal. Inside ADF, created a pipeline for data movement.
2. Created Linked Services for Azure Blob Storage and Azure Data Lake Storage Gen2 (even though they are technically the same storage account, we treat them according to our organizational logic within ADF).
3. Created datasets representing source data in the raw-data container and the destination in ADLS Gen2 where we want our hierarchical data stored.
4. Used a Copy Data activity within our pipeline to copy data from the Blob Storage dataset to the ADLS Gen2 dataset.



Why to do this step : This maintains a clear distinction between the untouched raw data and the copy of raw data ready for processing, facilitating better data management and governance.

**Step 4 : Transform Data Using Azure Databricks**

1. Setup Azure Databricks workspace and cluster created to run computations.
2. Mounted ADLS Gen2 to Azure Databricks by creating Application and keys for authentication.
3. Granted the permission to access and modify data in ADLS Gen2 storage account by assigning IAM role.
4. Now, by using Spark and Databricks the following transformations are done:

I. Filtered out thorough where the value in the "Venue" column equals "Venue," effectively removing any duplicate header rows.

II. Used regexp\_replace to remove annotations (e.g., [1], [2], [3]) and commas from the "Capacity" column.

III. Dropped the "Image" column

IV. Replaced null values in the "Tenant" column with the string "Unknown", ensuring that there are no missing values in this column.

V. Converted the "Capacity" column to an integer type.

VI. wrote the transformed data back to Azure Data Lake Storage Gen2

**Step 5 : EDA and Analysis in Azure Synapse**

1. Setted up a new Synapse workspace.
2. Connected Azure Data Lake Storage Gen2 to the Synapse workspace
3. Used the Synapse Studio to explore transformed data in ADLS Gen2.

Queries and Insights:

Query 1: Ranking Venues by Capacity within Each Country

SQL query ranks sports venues within each country by their capacity in descending order. The DENSE\_RANK() function is used to assign a rank without gaps in the ranking sequence for venues with the same capacity. It selects the country, city, venue, capacity, and calculated rank for each venue.

Insights:

1. Football is really popular, but other sports like car racing and horse racing also have big places for fans to watch the games.
2. Big cities are where most of the sports action happens.

Query 2 : Average Capacity by Sport and Country

Query calculates the average capacity and total number of venues for each sport within each country from the sports\_venue table. It groups the results by sport and country, then selects the sport, country, average capacity of venues, and count of venues in each group.

Insight:

There's a notable variation in the average capacity of venues for different sports within the same country, which suggests that some sports typically attract larger audiences and require bigger venues. The capacity also varies widely based on the sport and country combination, indicating that the popularity and cultural significance of sports can significantly influence the size of venues built to accommodate fans.

Query 3 : Relationship between venue capacity and the number of venues a sport has in each country.

This query calculates the average and total capacities for venues of each sport within each country, counts the number of venues per sport per country, and introduces a "Capacity\_Venue\_Score" which is the product of the average capacity and the number of venues. This score could help us understand not just where sports are popular, but where significant investments have been made into sports infrastructure (considering both the number of venues and their sizes).

Insights:

1. The significant "Capacity\_Venue\_Score" for American football in the US suggests a massive investment in sports infrastructure, with a high number of venues and large average capacities. This reflects the sport's popularity and its central role in American sports culture.
2. Motor racing in the US and association football (soccer) in countries like China and Brazil also show high scores, indicating substantial investments in these sports' infrastructures. These investments are not just in the number of venues but also in the scale of these venues, highlighting the global appeal and economic importance of motor racing and football. ​​

Query 4: Diversity of Sports Venue Sizes within Countries

This query calculates the average capacity of venues for each sport within each country, the range of capacities (difference between the largest and smallest venues), and the standard deviation of capacities (a measure of how spread out the capacities are). The standard deviation and range will give us an idea of whether a sport tends to have venues of similar sizes (low standard deviation and range) or if there's a wide variety in the sizes of venues for the sport (high standard deviation and range). The order by Capacity\_StdDev descending will show us where the greatest variability in venue sizes exists.

Insights:

1. The high standard deviation and capacity range for motor racing venues in France and Australia suggest a significant variability in venue sizes. This indicates that some motor racing events are hosted in vastly larger venues than others, potentially reflecting on the diverse scale of events (from local races to major international competitions).
2. Similarly, horse racing in Japan, Australia, and the US shows a wide range in venue capacities and a substantial standard deviation. This variability could reflect a mix of historic, smaller venues alongside modern, large-capacity facilities, illustrating the sport's evolution and its different scales of events within these countries. ​​

**Step 6 : Visualize the insights using Power BI.**

1. Connected the Data from the azure synapse to Power BI using SQL serverless endpoint.
2. Created a dashboard to visualize the insights.
3. The dashboard visualizations including :

Venue Capacity Table:

A table lists sports venues in the selected country, showcasing city, venue names, and their capacities. This provides a detailed breakdown of individual venues, which can be sorted and navigated through interactively.

Average Capacity by Sport and Country Bar Chart:

Adjacent to the table, a bar chart presents the average capacity of venues by sport within the selected country, giving insights into the popularity and scale of sporting infrastructure for different sports.

Scatter Plot - Venue Analysis by Sport:

Below the table, a scatter plot visualizes the relationship between the average venue capacity and the number of venues for different sports. The size of each circle corresponds to the "Capacity\_Venue\_Score", illustrating the combined factor of venue count and size.

Capacity Range and Standard Deviation Bar Chart:

To the right of the scatter plot, a horizontal bar chart displays the capacity range and standard deviation of venue sizes within the selected country, indicating the diversity in venue capacities for the chosen sport.