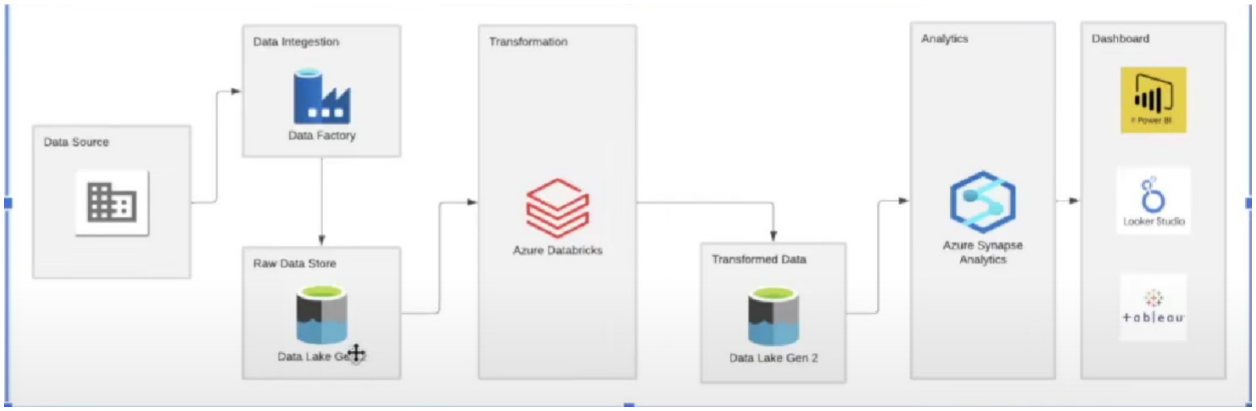
1. Project Goal

The primary objective of this data engineering project is to systematically traverse through the key phases of the Extract, Transform, Load (ETL) process and culminate in the development of an insightful dashboard. The project leverages various applications, including Azure Storage Account, Azure Data Factory, Databricks, Synapse Analytics, and Power BI, to accomplish this goal.

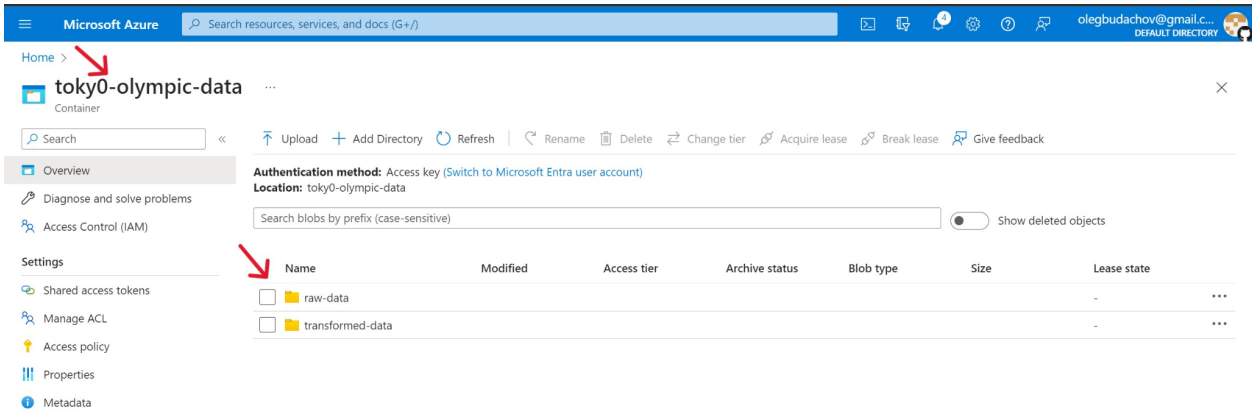


2. Data Source

Details

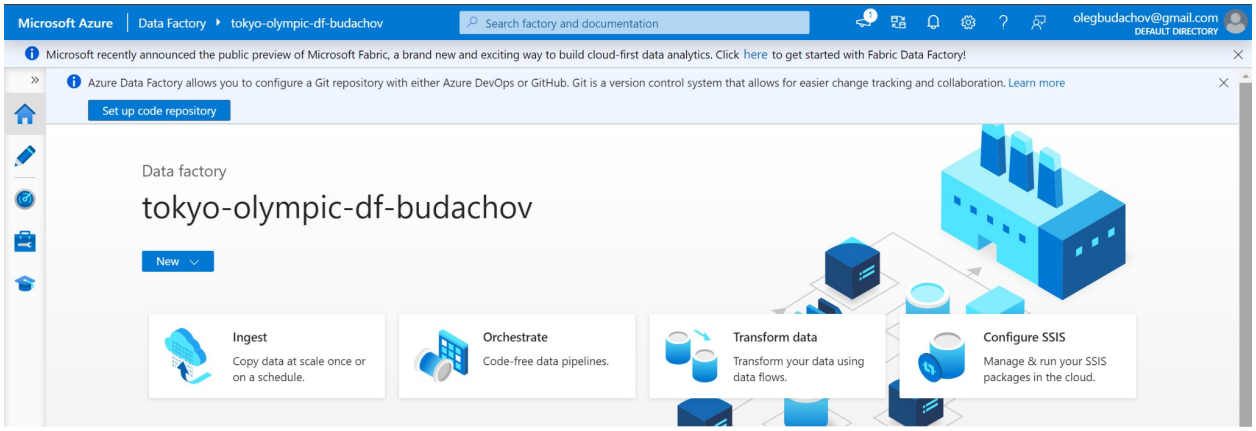
This contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics. This dataset contains the details of the Athletes, Coaches, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches and medals(gold, silver, bronze).
Source: Tokyo Olympics 2020 Website

3. Creating Azure storage account

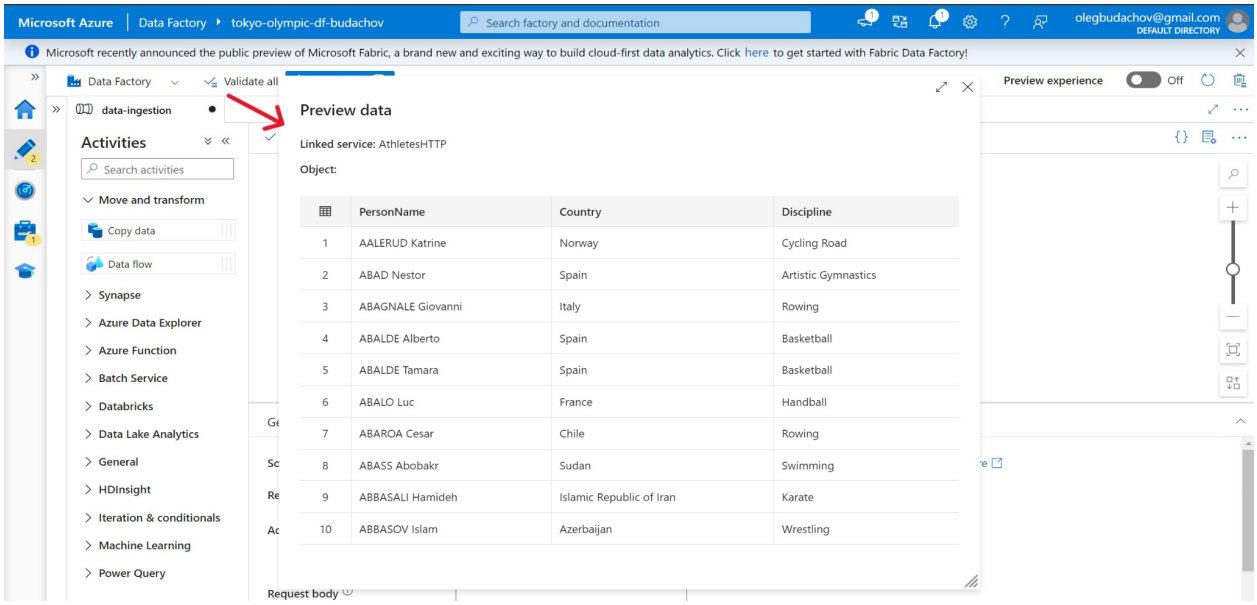


Within the container, we have established separate divisions: one designated for 'raw data' and the other for 'transformed data' to facilitate organization and accessibility.

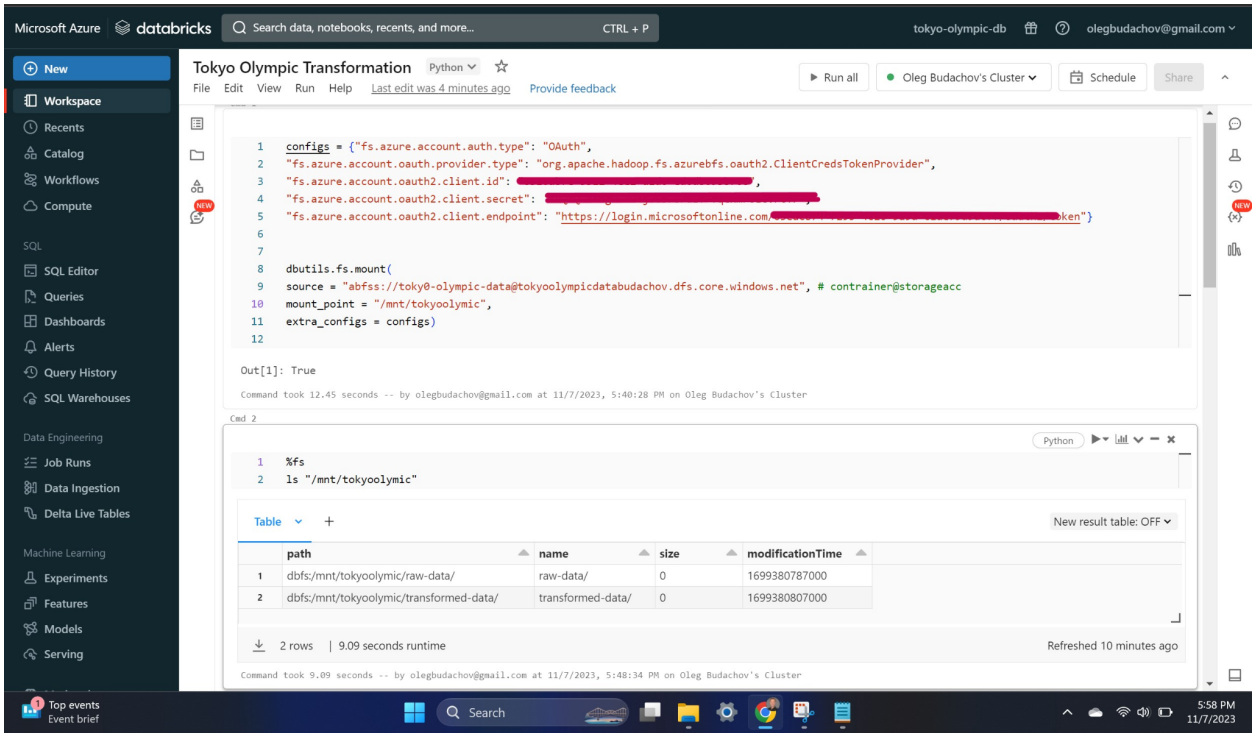
4. Using Azure data factory for importing data source to the Data Lake Storage

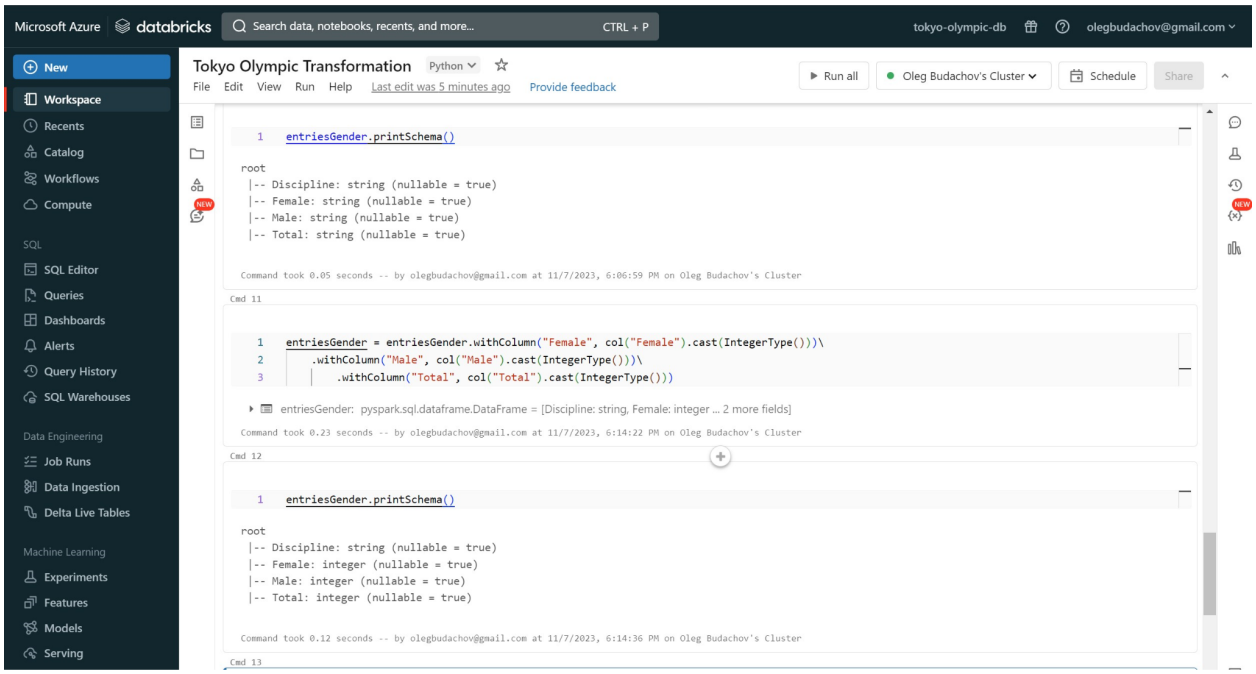


Verifying the successful upload of data within Data Factory



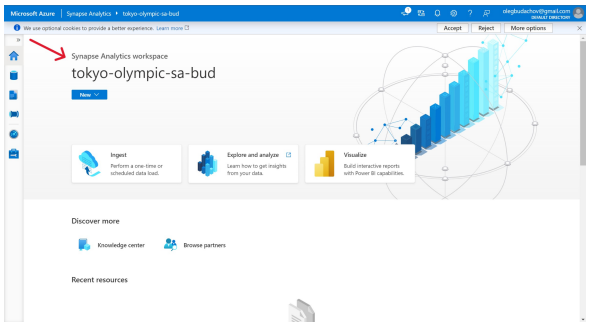
5. Leveraging Databricks for exploratory analysis and data transformation processes

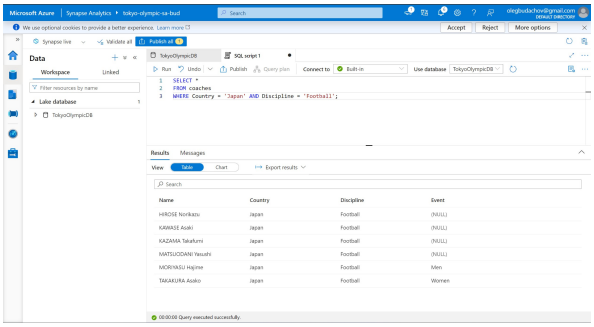


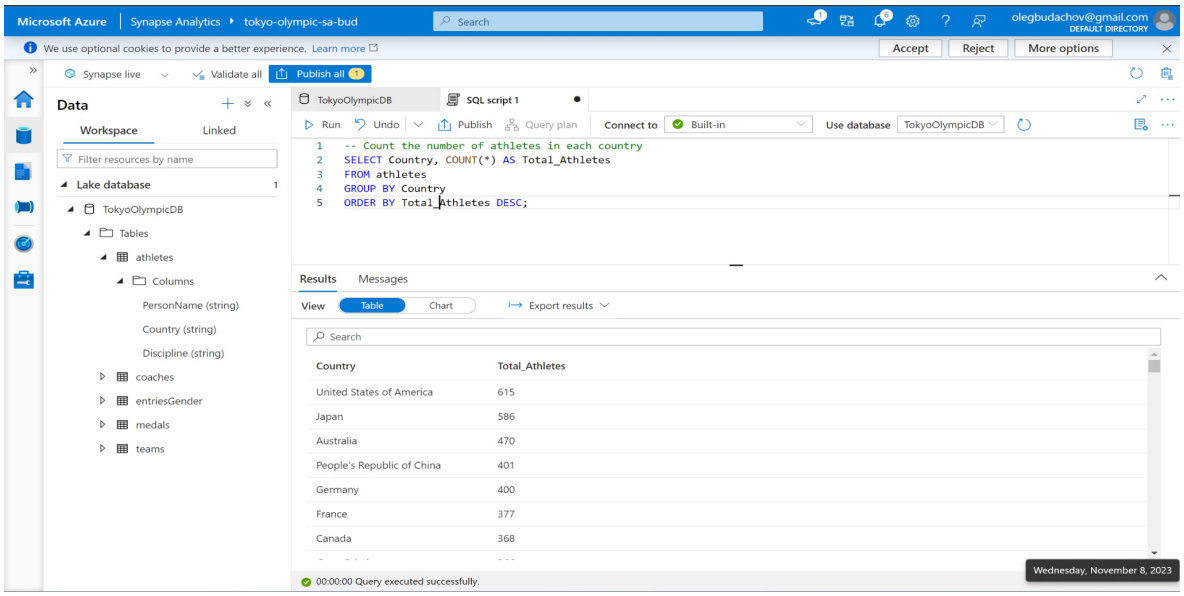


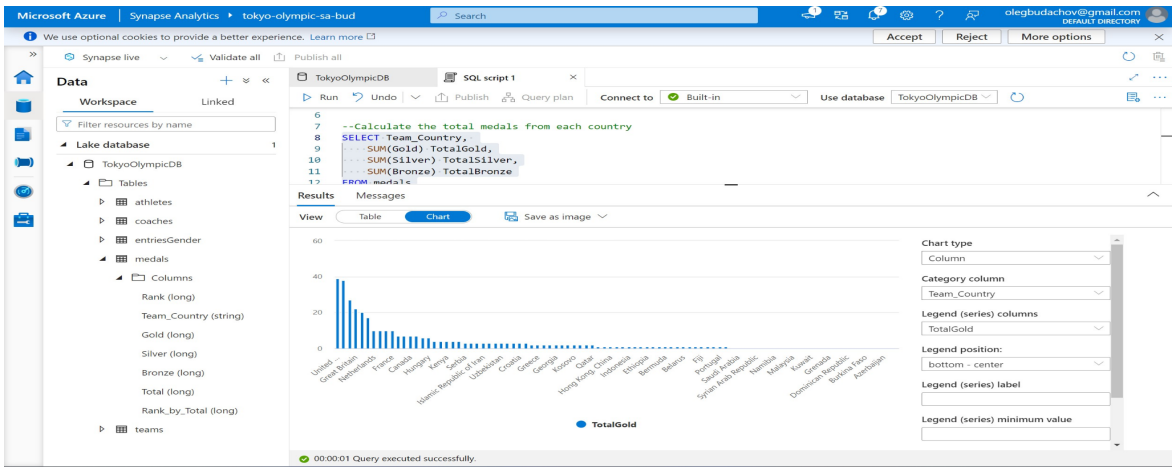
6. Synapse Analytics

We utilize Synapse Analytics after the data has been transformed through Databricks and uploaded to the transformed data storage for its comprehensive analytical capabilities. Synapse Analytics offers advanced querying, data warehousing, and scalable processing, enabling in-depth analytics, performance optimization, and seamless integration of transformed data for further insights and reporting.



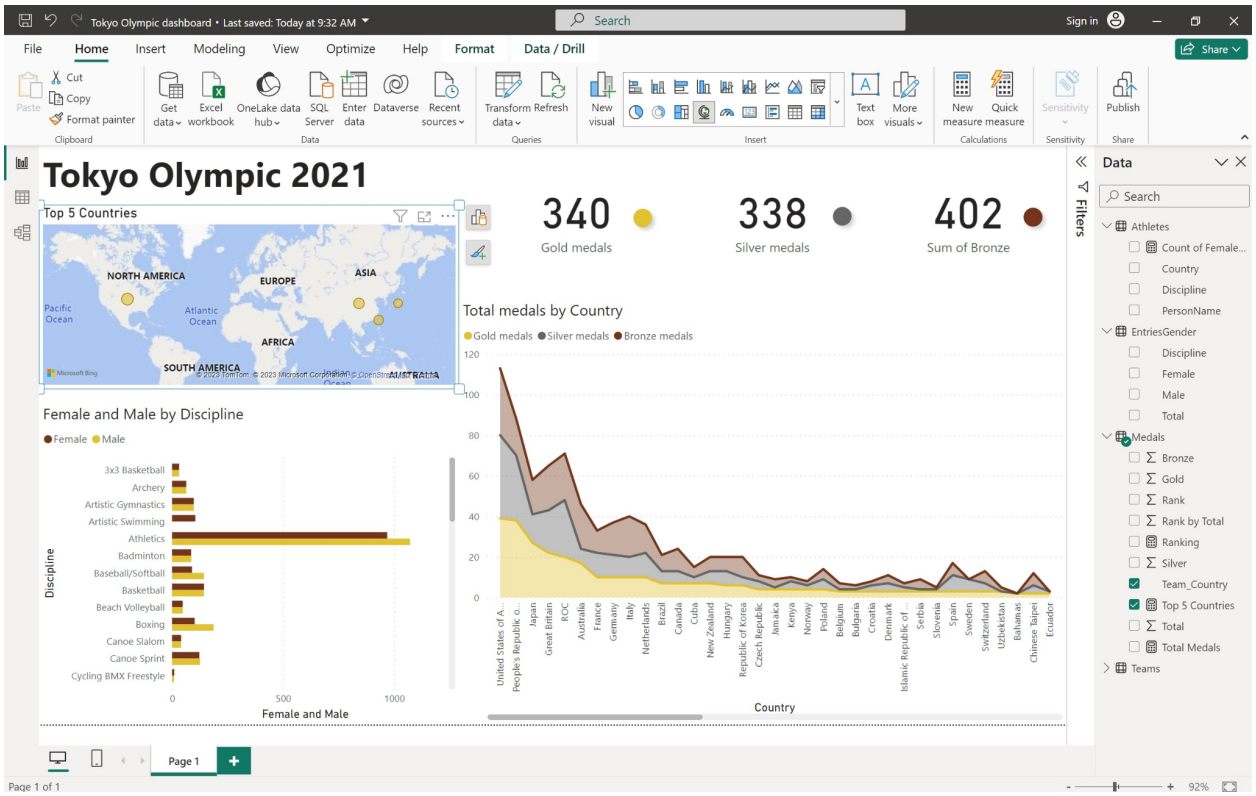






7. Creating Power BI Dashboard

We meticulously shape the data in Power Query Editor within Power BI by executing various tasks: splitting and removing old columns, crafting new columns, adjusting data types, establishing a data model with appropriate relationships. Subsequently, we design a dashboard showcasing key highlights concerning the Tokyo Olympic 2021, offering a comprehensive overview of pertinent information.



8. Conclusion

The project culminated in an extensive data transformation process involving various stages. Initially, the data underwent cleansing, reformatting, and structuring within Databricks, followed by uploading the transformed data to storage facilitated by Data Factory. The curated dataset was then further refined and shaped using Power Query Editor within Power BI. Throughout these stages, the old columns were modified, new columns were created, data types were adjusted, and a robust data model was established, ensuring appropriate relationships for comprehensive insights. This diligent process paved the way for the creation of an informative and visually compelling dashboard, focusing on the major highlights of the Tokyo Olympic 2021.