**Knowledge is learning new something every day. Wisdom is letting go of some bad habits everyday.**

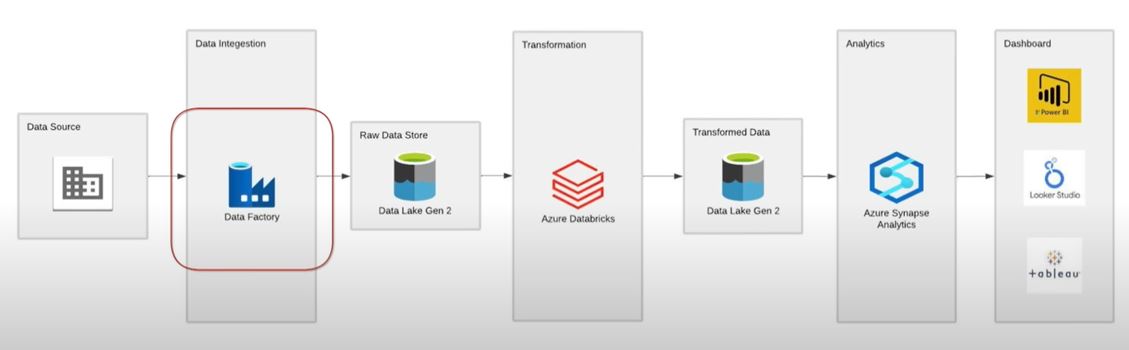
I am presenting a simple data engineering project which I explored from one of you tube video.

**Here we are taking sample data set from kaggle -** [**https://www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo**](https://www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo)

**Note:** For learning purpose we are trying to use different Azure services.

* **Azure Data Factory:** Azure data factory is data integration service that enables you to create, schedule, and manage data pipelines for efficient data movement and transformation between various sources and destinations in azure and beyond. It simplifies ETL(Extract, Transform, Load) and data integration tasks.
* **Azure Data Lake Gen2:** Data lake solution that combines the capabilities of a data lake with the power of azure blob storage, allowing you to store and analyze large volumes of structured and unstructured data with enhanced performance, security and analytics capabilities.
* **Azure Databricks:** Databricks is a unified analytics platform built on top of apache spark, designed to help data engineers and data scientists collaborate on big data processing and machine learning tasks. It provides tools for data exploration, data processing and building machine learning models in a collaborative and scalable environment.
* **Azure Synapse Analytics:** SQL Data warehouse is a cloud-based analytics service provided by Microsoft azure. It combines big data and data warehousing into a single integrated platform, allowing organizations to analyze and process large volume of data for business intelligence and data analytics purposes.

**The image shows the end-to-end architecture /pipeline of project.**



1. We are taking data from data source. We will extract data from API using Azure Data factory and store this data as raw data folder in Azure data lake gen 2.
2. We are using azure databricks service to do some transformation on data.
3. Later we will store transformed data back to Azure data lake gen 2 in transformed data folder. Later this transformed data is used to do business analytics.
4. For analytics we are using azure synapse analytics services. And final we can present analytics results in any kind of dashboard tools such as tableau.