END TO END DATA ENGINEERING PROJECT – TOKYO OLYMPICS 2021

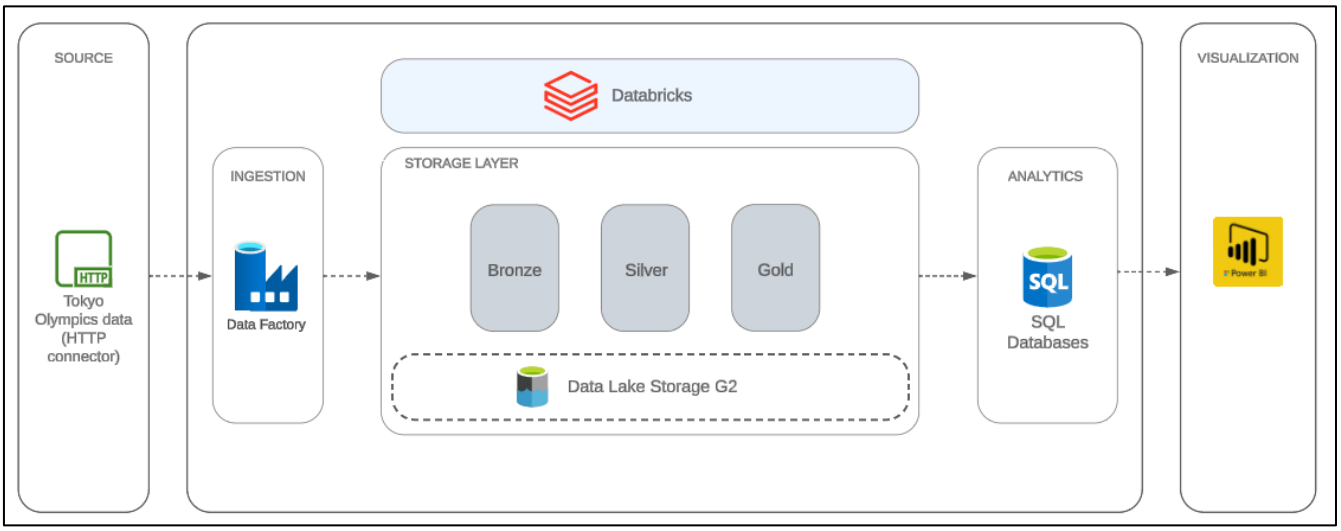
DESCRIPCIÓN DEL PROYECTO

En este proyecto se analizará los datos olímpicos que se encuentran disponibles en kaggle construyendo un proyecto end-to-end, donde se utilizará los principales servicios de la nube de Azure.

La data será ingestada desde un servicio **http** (GitHub) utilizando **Azure Data Factory y** almacenada en su estado original en la capa **Bronze** en un **Data Lake Storage G2**, y utilizando el servicio de Azure Databricks será transformada pasando de la capa **Bronze** a la capa **Silver** y posteriormente a la capa **Gold.**

Se utilizará **SQL Database** para realizar consultas a la data limpia y para que la data pueda ser leída por herramientas como **Power Bi** para la visualización y presentación de reportes.

ARQUITECTURA DEL PROYECTO



DEFINICIÓN DE SERVICIOS

AZURE DATA FACTORY

Servicio de integración de datos basados en la nube, y sirve para orquestar y automatizar el movimiento y transformación de datos.

AZURE DATA LAKE STORAGE GEN2

Data Lake Storage Gen2 converge las capacidades de **Azure Data Lake Storage Gen1** con **Azure Blob Storage**.

AZURE DATABRICKS

Plataforma que proporciona herramientas para la exploración y el procesamiento de datos y la creación de modelos de aprendizaje automático en entornos colaborativos y escalables.

SQL DATABASE

SQL Database es un servicio de base de datos en la nube.

VISUALIZACIÓN DE REPORTES