# **Day 14 - Medallion Architecture in Azure**

**Date:** 26/06/2025

### **1️⃣ Understanding the Medallion Architecture**

* The Medallion Architecture consists of three layers:  
  + **Bronze Layer**: Raw data ingestion from Azure Blob Storage.
  + **Silver Layer**: Cleaned and transformed data for analysis.
  + **Gold Layer**: Aggregated and business-insight-ready data.

### **2️⃣ Setting Up Azure Environment**

* **Azure Blob Storage Account** was created.
* Containers named bronze, silver, and gold were added under **Blob Service**.
* **Access keys** were retrieved to allow secure mounting in Databricks.

### **3️⃣ Deploying Azure Databricks**

* Created a **Databricks workspace** from Azure.
* Launched the workspace and deployed a cluster (pima-cluster) with default configuration.

### **4️⃣ Mounting Blob Containers to Databricks**

Using dbutils.fs.mount, each container (bronze, silver, and gold) was mounted to:

* /mnt/bronze
* /mnt/silver
* /mnt/gold

Example mount command used:

dbutils.fs.mount(

source = f"wasbs://bronze@<storage\_account>.blob.core.windows.net/",

mount\_point = "/mnt/bronze",

extra\_configs = {"fs.azure.account.key.<storage\_account>.blob.core.windows.net": "<access\_key>"}

)

### **5️⃣ Bronze Layer Implementation**

* **Raw CSV** (PIMA Diabetes dataset) was uploaded to the bronze container.
* Data was read with a custom schema and stored as a **Delta Table** in the bronze path:

raw\_df.write.format("delta").mode("overwrite").save("/mnt/bronze/delta/pima\_raw")

### **6️⃣ Silver Layer Implementation**

* Read data from Bronze Delta table.
* Replaced zero values in critical columns (e.g., Glucose, BloodPressure, BMI) with nulls or median.
* Derived new columns like **BMI Category**.
* Transformed DataFrame saved to:

/mnt/silver/delta/pima\_cleaned

### **7️⃣ Gold Layer Implementation**

* Loaded cleaned data from Silver.
* Performed group-wise aggregations (e.g., average BMI by Outcome).
* Final dataset written to Gold path:

/mnt/gold/delta/pima\_summary

### **8️⃣ Key Learnings**

* Delta Lake makes data pipelines more efficient and queryable with version control.
* Mounting blob storage allows easy integration between Azure services.
* Data engineering pipelines follow a structured transformation path from raw to insight.

### **9️⃣ Additional Insights**

* Optionally, a **Platinum Layer** can be added using Power BI or Databricks SQL for visualization and trend analysis.
* This project lays the groundwork for machine learning model integration on clean, structured datasets.