# Azure Clinical Data Warehouse – Project Journal

> This document tracks all the steps, decisions, and work done for the \*\*Azure Clinical Data Warehouse pipeline portfolio project\*\*.

## 1. Project Overview

Goal: Build a production-like end-to-end clinical data warehouse pipeline on Azure using simulated data.

Pipeline Architecture: Ingest → Bronze / Raw zone → Silver / Cleaned → Gold / Curated → Analytics / Power BI → CDS demo

Tools & Services:

- Databases: Postgres, SQL Server, MongoDB  
- Data Lake: ADLS Gen2  
- ETL / Transform: Azure Data Factory, Databricks  
- Analytics: Synapse / Power BI  
- Governance / Security: Azure Purview, Key Vault, RBAC

## 2. Step 1 – Data Simulation

Actions Taken:

- Generated synthetic data using Python + Faker.

- Created 5 datasets:  
 \* patients.csv – Postgres  
 \* encounters.csv – Postgres  
 \* labs.csv – Postgres  
 \* registry.csv – SQL Server  
 \* imaging.csv – SQL Server  
 \* device\_readings.csv – MongoDB

## 3. Step 2 – Database Setup

Actions Taken:

- Imported patients, encounters, labs into Postgres (via DBeaver).  
- Imported registry and imaging into SQL Server (via DBeaver).  
- Inserted device\_readings into MongoDB (via mongoimport).

## 4. Step 3 – Azure Platform Setup

Actions Taken:

- Created Resource Group: rg-azure-clinical-dw-dev  
- Provisioned ADLS Gen2 with Bronze, Silver, Gold containers  
- Provisioned Azure Data Factory in same region  
- Installed Self-hosted Integration Runtime on local machine  
- Configured Linked Services:  
 \* Postgres (patients, encounters, labs)  
 \* SQL Server (registry, imaging)  
 \* MongoDB (device\_readings)  
- Verified connectivity to all sources

## 5. Step 4 – Data Ingestion

Actions Taken:

- Built pipelines in ADF to copy Postgres, SQL Server, MongoDB data into ADLS Bronze  
- Verified row counts after copy  
- Performed light cleaning (trim spaces, date format standardization)  
- Stored raw CSV and JSON in Bronze as-is

## 6. Step 5 – Transformation

Actions Taken:

- Provisioned Azure Databricks  
- Connected to ADLS Gen2  
- Ran first test notebook to move data from Bronze → Silver  
- Applied schema alignment and type casting

## 7. Step 6 – Synapse & Power BI

Actions Taken:

- Created Synapse Workspace and linked to ADLS  
- Tested serverless SQL pool queries on Silver data  
- Connected Power BI to Synapse SQL endpoint  
- Built initial patient demographics dashboard

## 8. Step 7 – Governance & Security

Actions Taken:

- Configured RBAC for ADLS containers  
- Enabled encryption at rest and TLS  
- Set up Azure Key Vault for secrets  
- Planned to add Purview later for lineage tracking