Diabetes Risk Analysis ETL Pipeline

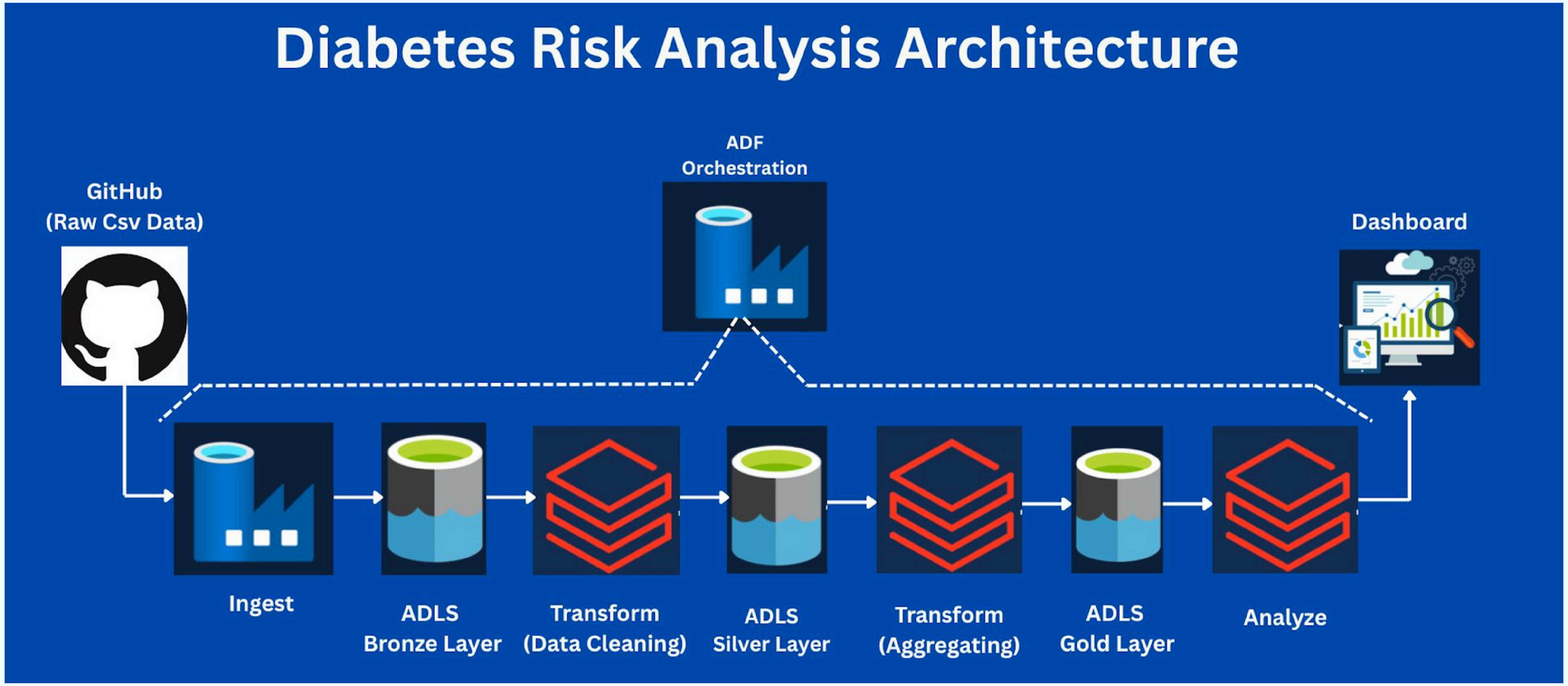
Using Azure Data Factory and Azure Databricks

PRESENTED BY

Puneeth Kumar Amudala

Project Introduction

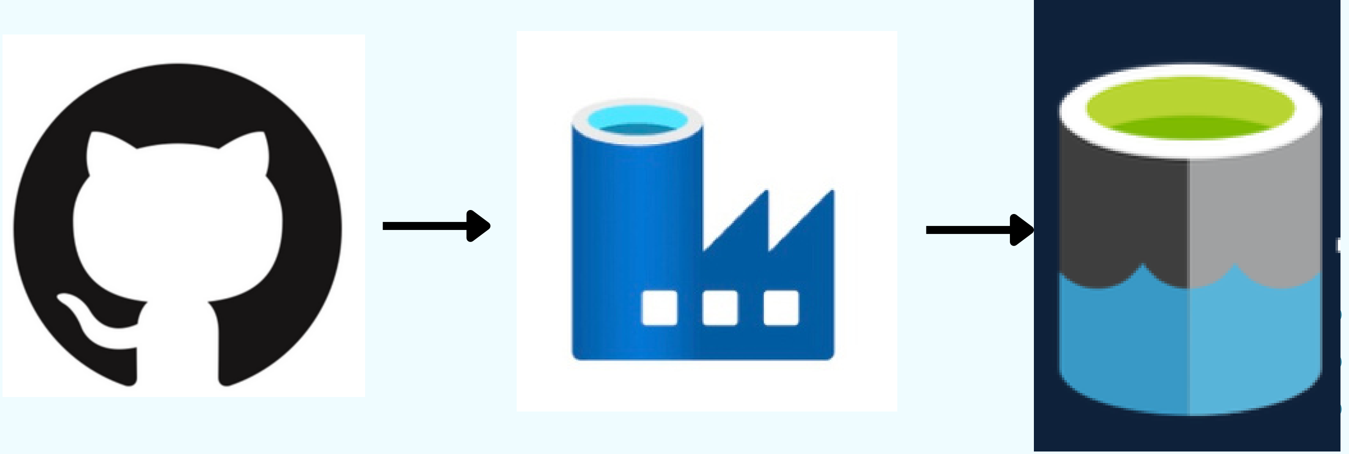
- DATASET FROM GITHUB (CSV)
- HEALTH INDICATORS RELATED TO DIABETES
- BATCH PROCESSING



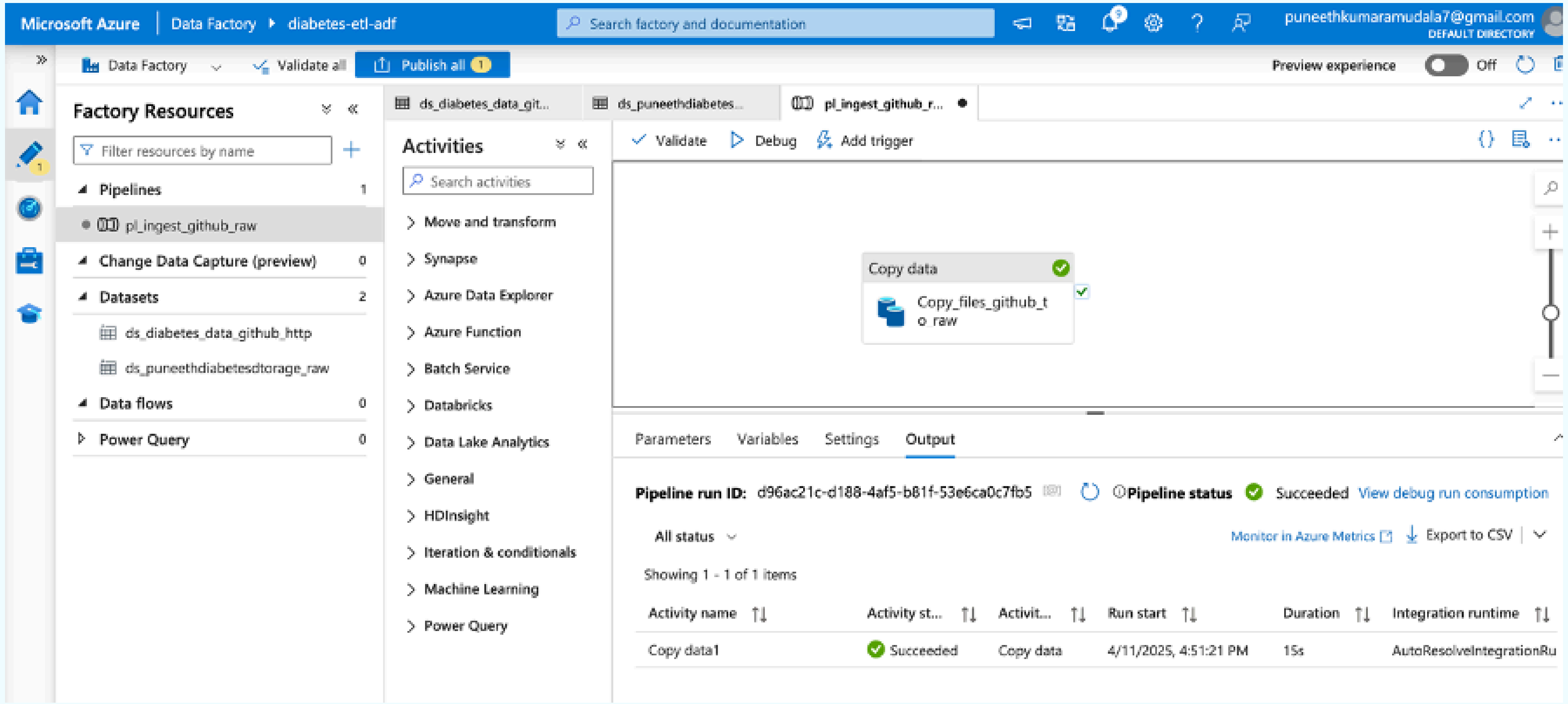
Diabetes Risk Analysis Architecture

Data Ingestion

Tool: Azure Data Factory


GITHUB → ADF → BRONZE LAYER

- PULLED FROM GITHUB USING HTTP CONNECTOR
- STORED RAW DATA IN ADLS BRONZE LAYER



Pipeline run ID: d96ac21c-d188-4af5-b81f-53e6ca0c7fb5 **Pipeline status** ✓ Succeeded View debug run consumption

All status

Monitor in Azure Metrics | Export to CSV

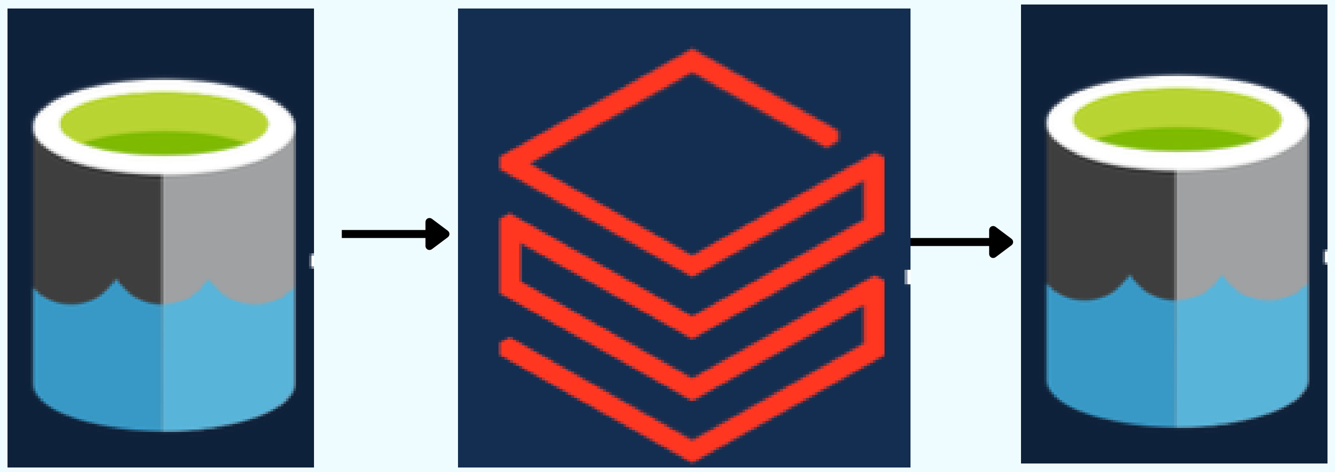
Showing 1 - 1 of 1 items

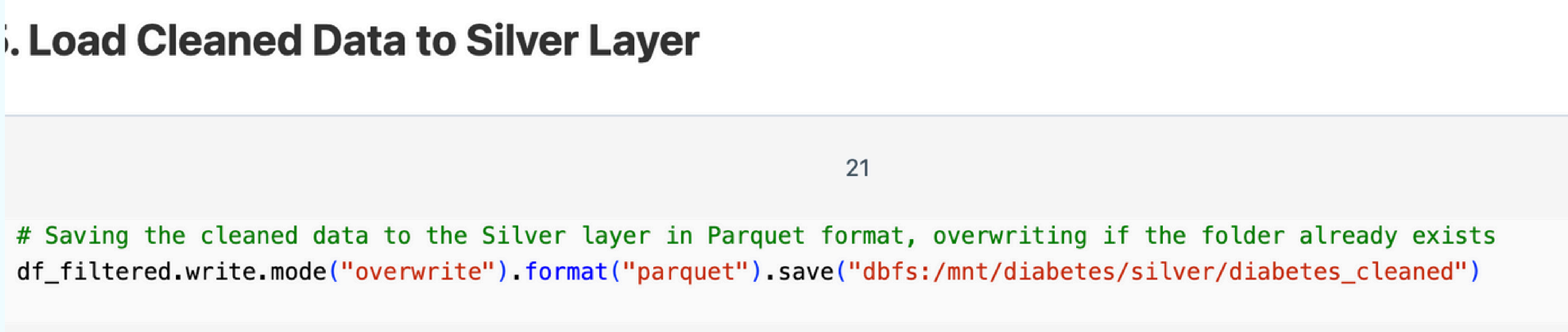
Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime
Copy data1	✓ Succeeded	Copy data	4/11/2025, 4:51:21 PM	15s	AutoResolveIntegrationRu

Data Transformation

Tool: Azure Databricks

- BRONZE TO SILVER
- APPLIED SCHEMA
- REMOVED NULLS, DUPLICATES
- SAVED AS PARQUET IN SILVER LAYER

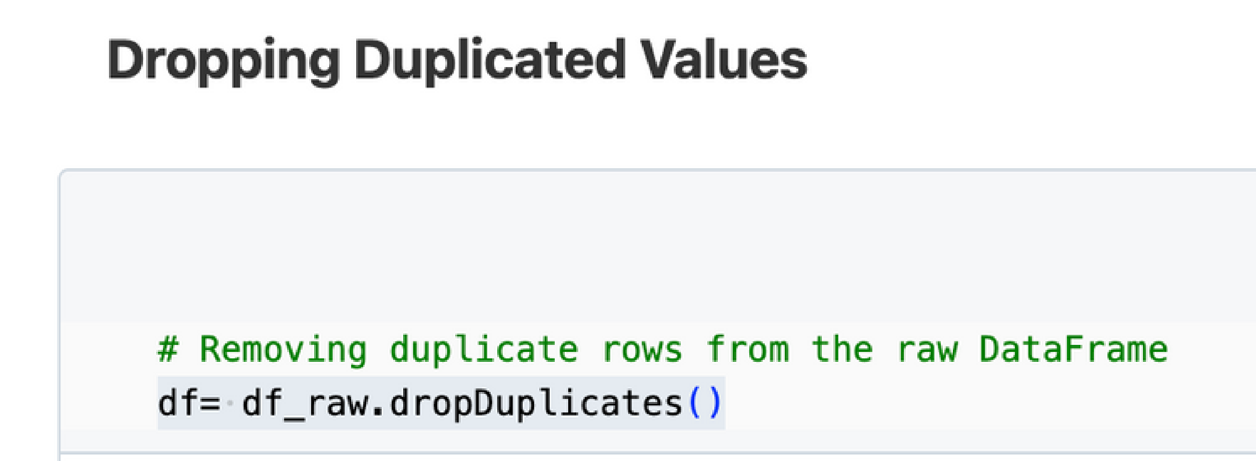




. Load Cleaned Data to Silver Layer

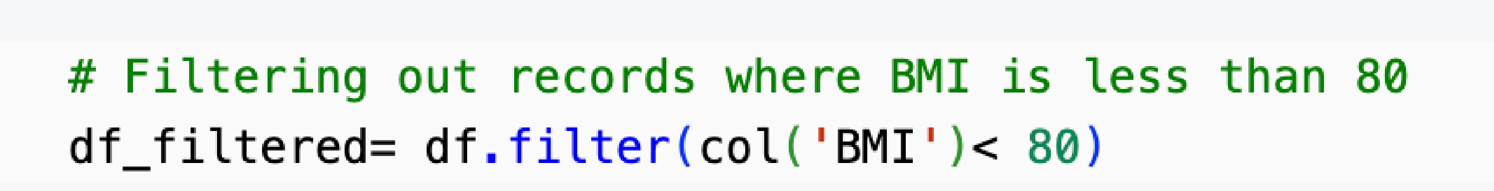
21

```
# Saving the cleaned data to the Silver layer in Parquet format, overwriting if the folder already exists
df_filtered.write.mode("overwrite").format("parquet").save("dbfs:/mnt/diabetes/silver/diabetes_cleaned")
```



Dropping Duplicated Values

```
# Removing duplicate rows from the raw DataFrame
df= df_raw.dropDuplicates()
```

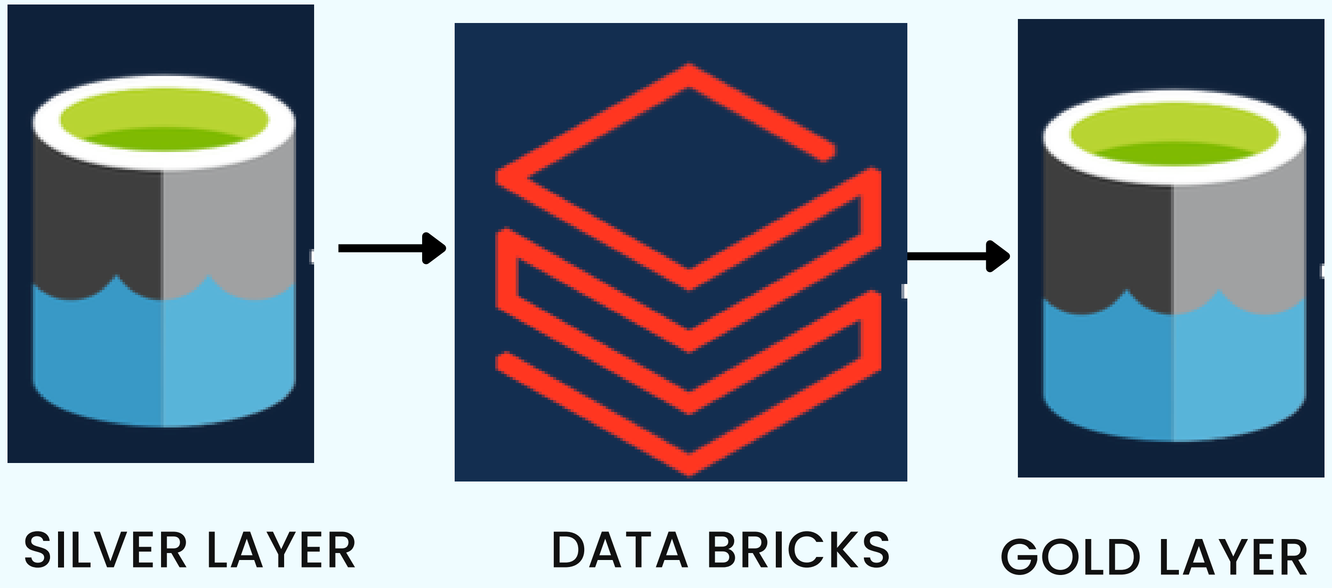


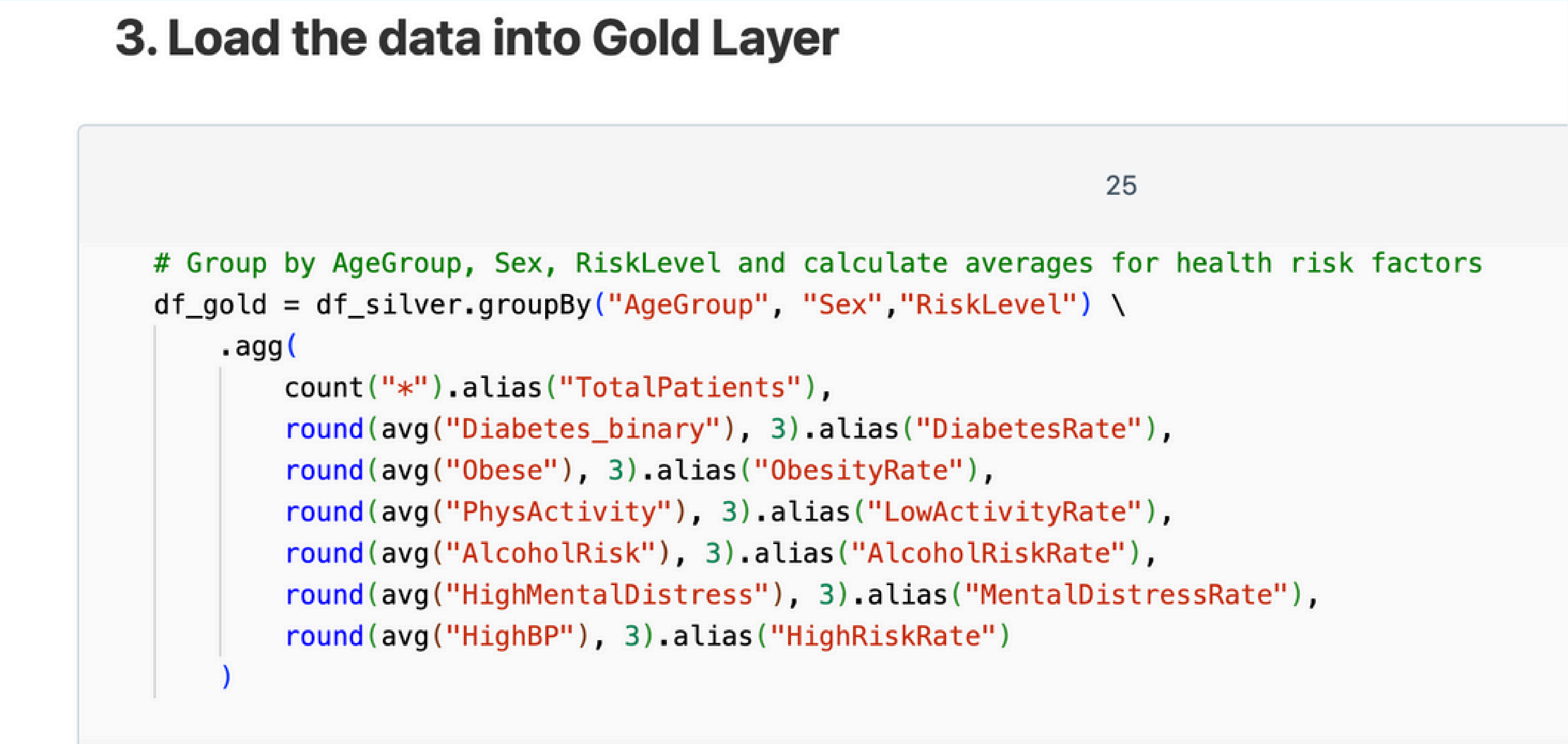
```
# Filtering out records where BMI is less than 80
df_filtered= df.filter(col('BMI')< 80)
```

Data Loading

Tool: Azure Databricks

- BRONZE TO SILVER
- FEATURE ENGINEERING
- AGGREGATIONS
- SAVED AS DELTA IN GOLD LAYER

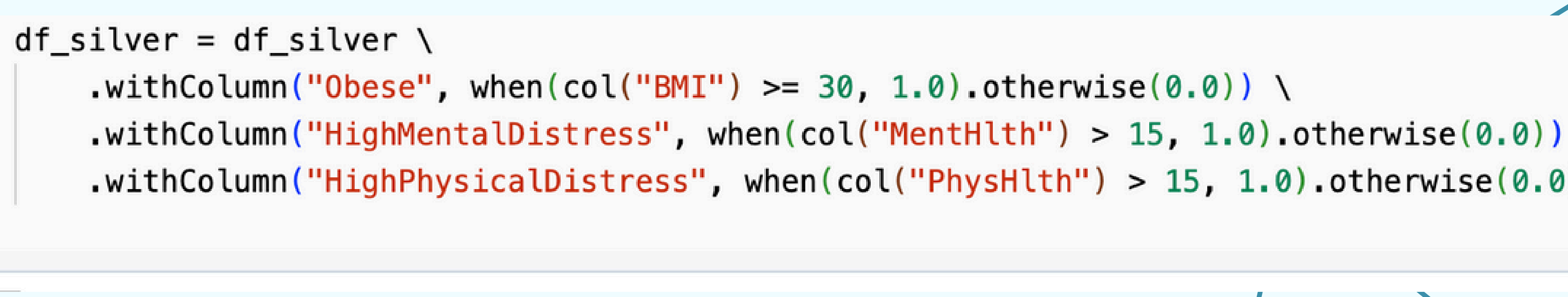




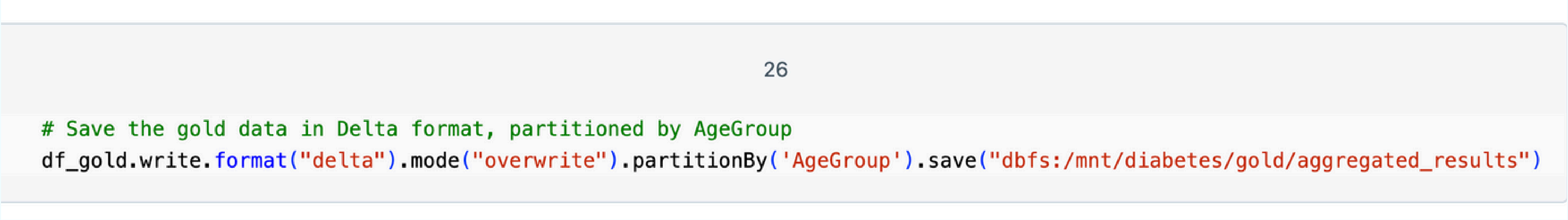
3. Load the data into Gold Layer

25

```
# Group by AgeGroup, Sex, RiskLevel and calculate averages for health risk factors
df_gold = df_silver.groupBy("AgeGroup", "Sex","RiskLevel") \
    .agg(
        count("*").alias("TotalPatients"),
        round(avg("Diabetes_binary"), 3).alias("DiabetesRate"),
        round(avg("Obese"), 3).alias("ObesityRate"),
        round(avg("PhysActivity"), 3).alias("LowActivityRate"),
        round(avg("AlcoholRisk"), 3).alias("AlcoholRiskRate"),
        round(avg("HighMentalDistress"), 3).alias("MentalDistressRate"),
        round(avg("HighBP"), 3).alias("HighRiskRate")
    )
```



```
df_silver = df_silver \
    .withColumn("Obese", when(col("BMI") >= 30, 1.0).otherwise(0.0)) \
    .withColumn("HighMentalDistress", when(col("MentHlth") > 15, 1.0).otherwise(0.0))
    .withColumn("HighPhysicalDistress", when(col("PhysHlth") > 15, 1.0).otherwise(0.0
```



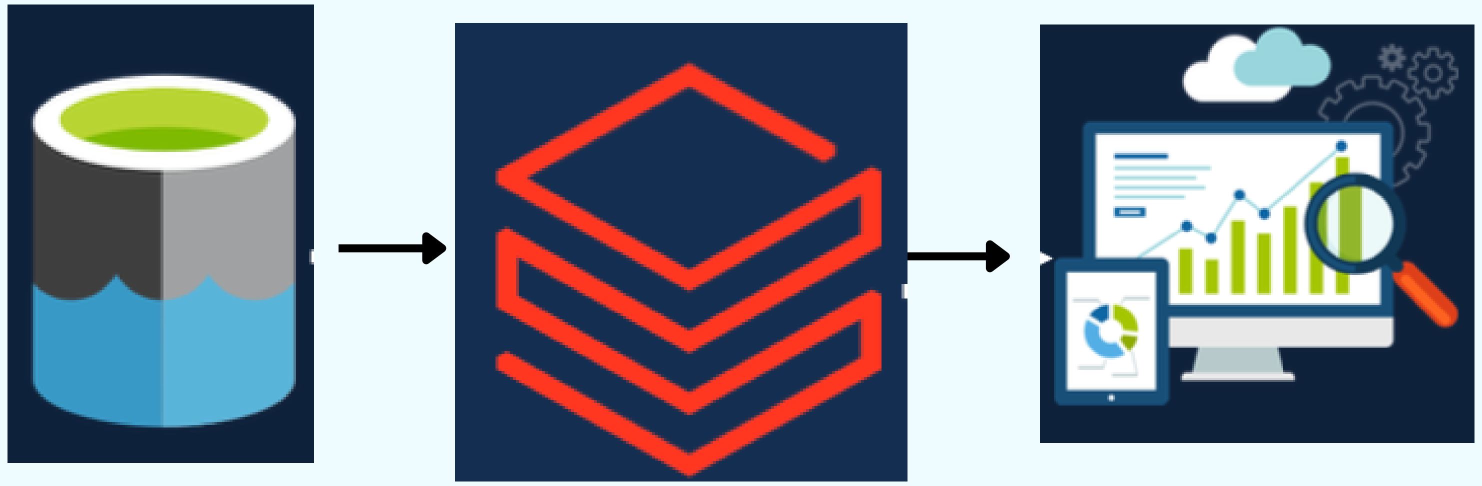
26

```
# Save the gold data in Delta format, partitioned by AgeGroup
df_gold.write.format("delta").mode("overwrite").partitionBy('AgeGroup').save("dbfs:/mnt/diabetes/gold/aggregated_results")
```

Data Analysis

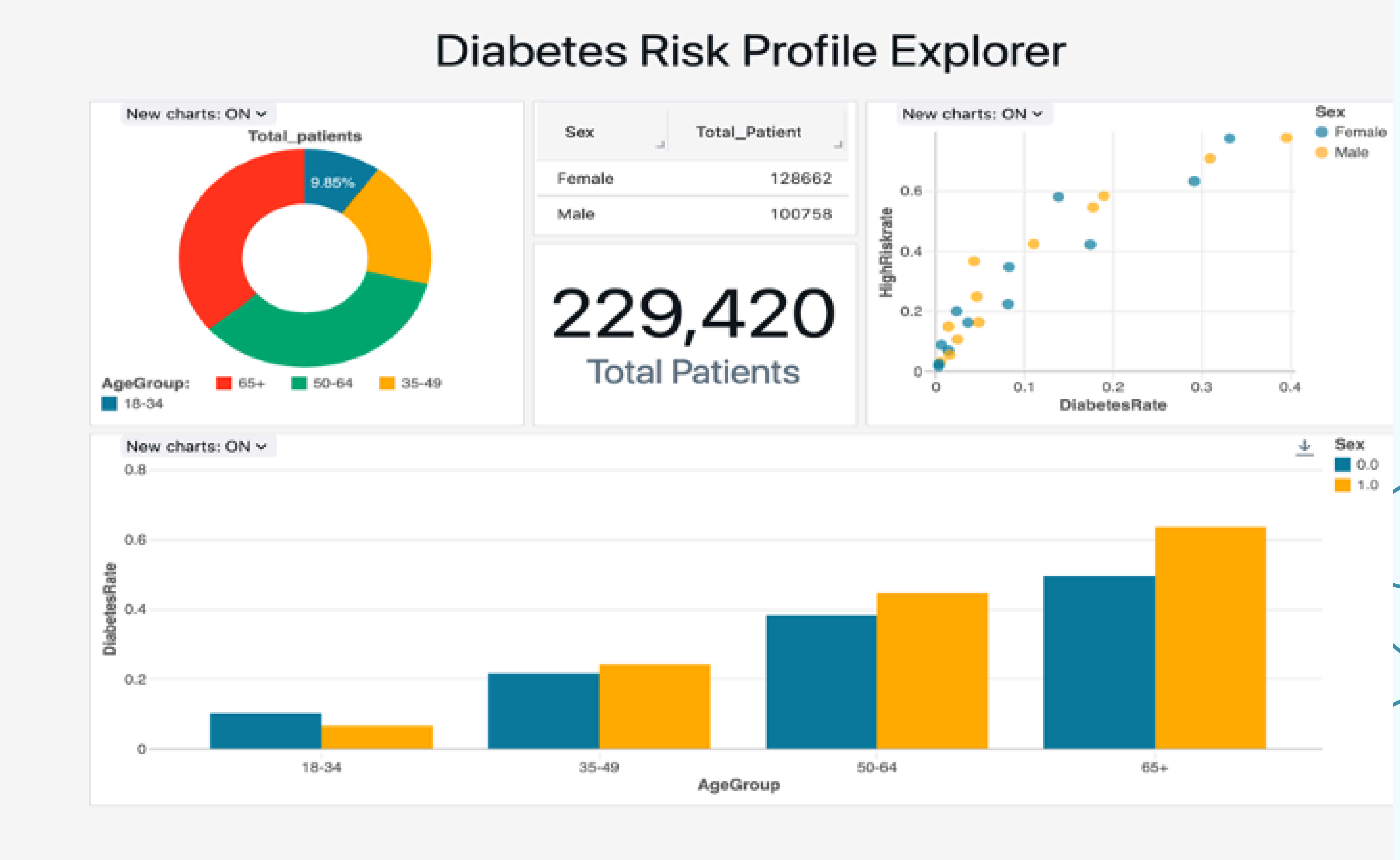
Tool: Azure Databricks

DASHBOARD-1



GOLD LAYER → DATA BRICKS → DASHBOARD

- **PIE CHART:** PATIENT DISTRIBUTION BY AGE GROUP
- **KPI BOX:** TOTAL PATIENTS BY GENDER
- **SCATTER PLOT:** DIABETESRATE VS HIGHRISKRATE BY GENDER
- **BAR CHART:** DIABETESRATE BY AGEGROUP AND SEX



Diabetes Risk Profile Explorer

Sex	Total_Patient
Female	128662
Male	100758

229,420

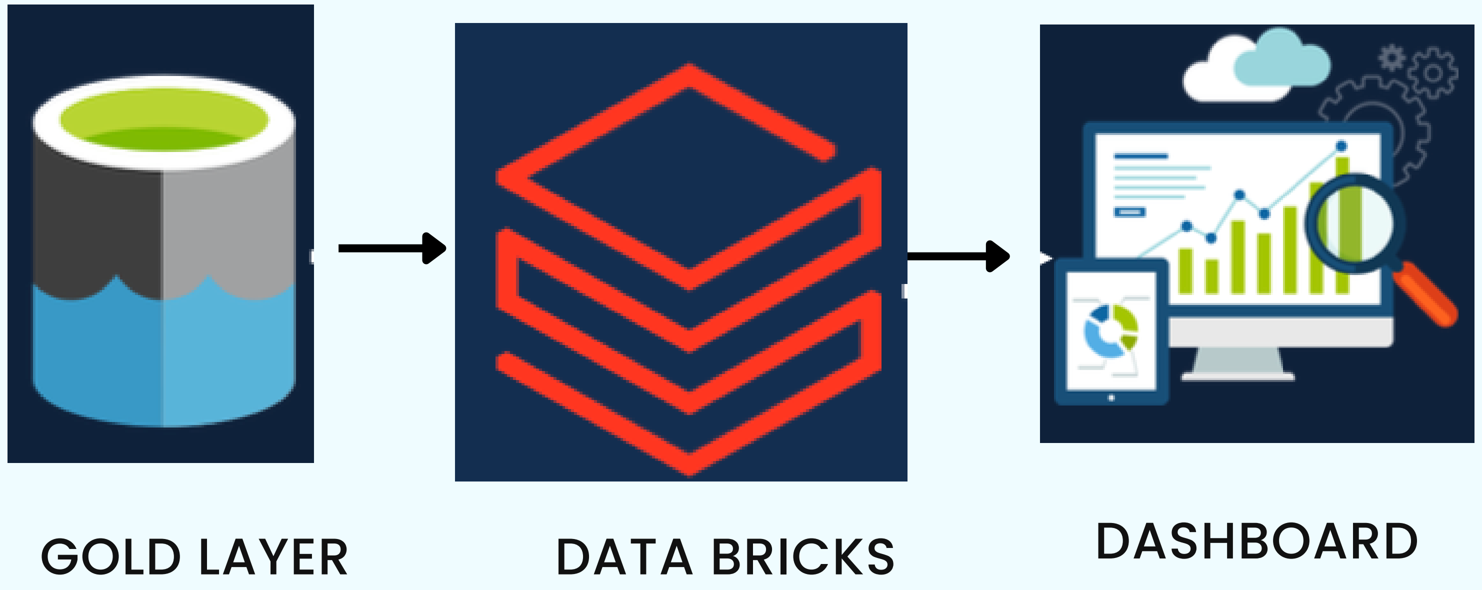
Total Patients

Data Analysis

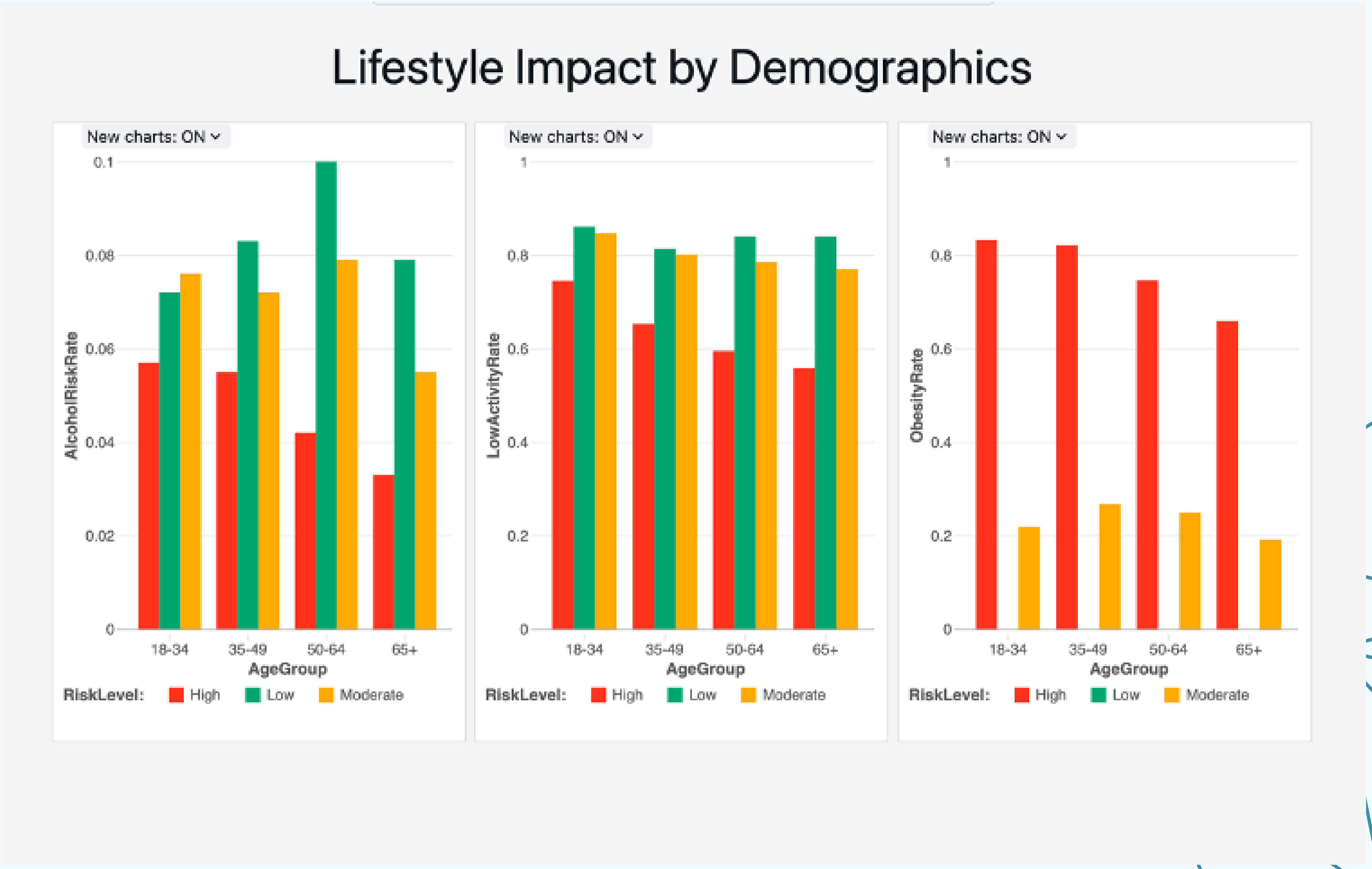
Tool: Azure Databricks

DASHBOARD-2

- **ALCOHOLRISKRATE CHART:** ALCOHOL HABITS BY AGEGROUP AND RISKLEVEL
- **LOWACTIVITYRATE CHART:** PHYSICAL INACTIVITY PATTERNS
- **OBESITYRATE CHART:** OBESITY TRENDS ACROSS SEGMENTS



GOLD LAYER → DATA BRICKS → DASHBOARD



Lifestyle Impact by Demographics

ETL Orchestration

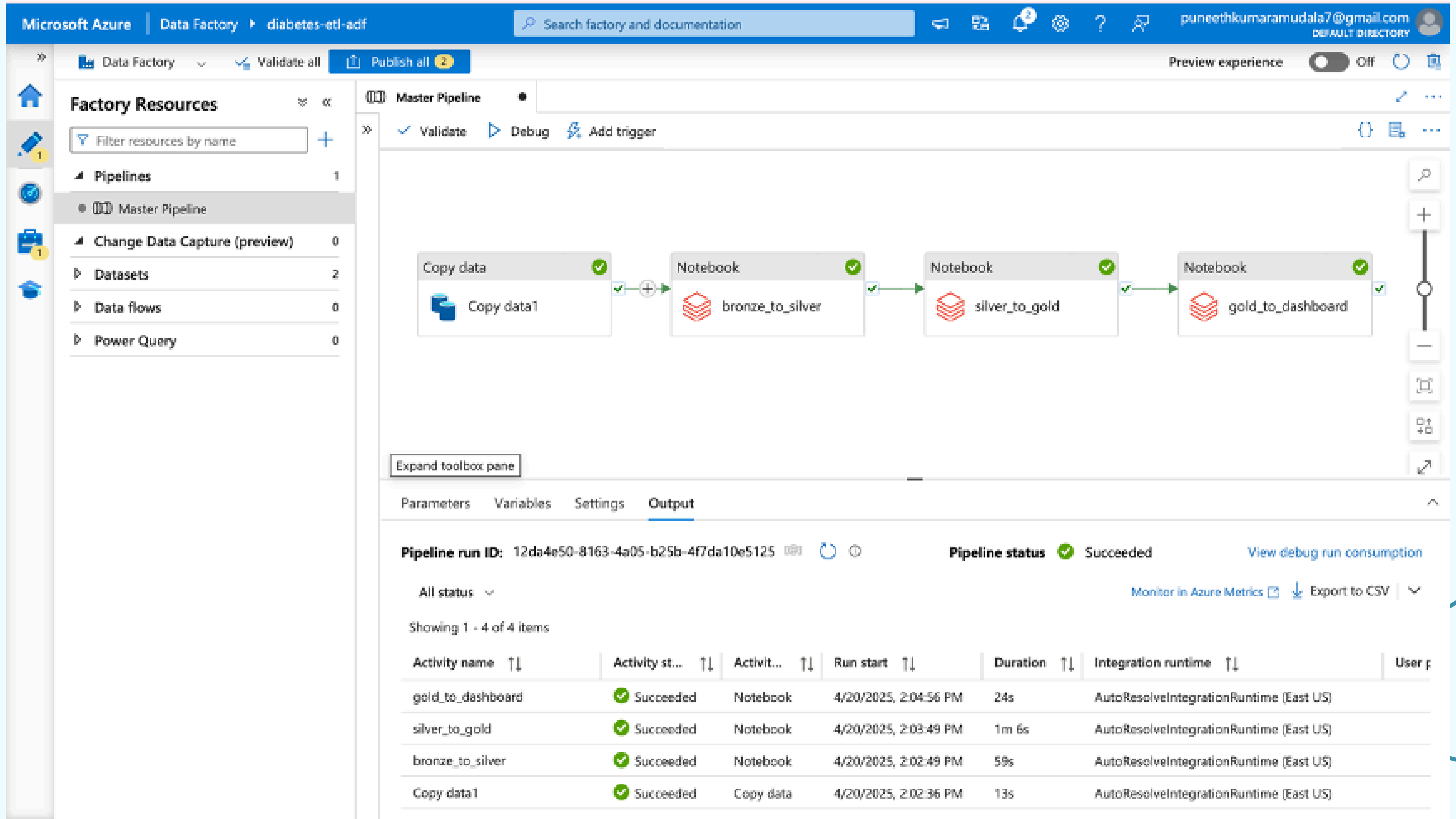
Tool: Azure Data Factory

1. COPY RAW DATA TO BRONZE

2. TRANSFORM BRONZE ➜ SILVER USING `BRONZE_TO_SILVER` NOTEBOOK

3. TRANSFORM SILVER ➜ GOLD USING `SILVER_TO_GOLD` NOTEBOOK

1. GENERATE VISUAL INSIGHTS USING `GOLD_TO_DASHBOARD` NOTEBOOK



Pipeline run ID: 12da4e50-8163-4a05-b25b-4f7da10e5125

Pipeline status ✓ Succeeded

Showing 1 - 4 of 4 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User p
gold_to_dashboard	✓ Succeeded	Notebook	4/20/2025, 2:04:56 PM	24s	AutoResolveIntegrationRuntime (East US)	
silver_to_gold	✓ Succeeded	Notebook	4/20/2025, 2:03:49 PM	1m 6s	AutoResolveIntegrationRuntime (East US)	
bronze_to_silver	✓ Succeeded	Notebook	4/20/2025, 2:02:49 PM	59s	AutoResolveIntegrationRuntime (East US)	
Copy data1	✓ Succeeded	Copy data	4/20/2025, 2:02:36 PM	13s	AutoResolveIntegrationRuntime (East US)	

Key Findings

DATA ANALYSIS

- Obesity strongly linked to risk
- Males aged 40–60 at higher risk

CHALLENGE FACED

Extracting data Directly from kaggle

SOLUTION

Downloaded the data and Uploaded in My Github

Conclusion

Project Summary

- Full ETL pipeline from ingestion to insights
- Used Medallion Architecture (Bronze → Gold)
- Learned orchestration, storage formats, and dashboards

Thank You