**Phase 2: Data Pipeline Development**

This stage transforms raw CSV survey data into clean, enriched datasets that are ready for analytics and dashboarding. The pipeline follows a multi-layer architecture:

A diagram of data quality

AI-generated content may be incorrect.

Reference: [Blog](https://www.advancinganalytics.co.uk/blog/medallion-architecture)   
  
**Phase-2**: Bronze ➜ Silver: Cleansing and structuring  
**Phase-3**: Silver ➜ Gold: Feature engineering and aggregations

## Bronze ➜ Silver

Input:  
**Extract**

bronze/diabetes.csv (raw BRFSS health records)

Key Steps:

**Transform**

* - Apply schema to enforce data types
* - Drop duplicates
* - Remove invalid entries (e.g., extreme BMI values)
* - Drop nulls in key health indicators

Output:

**Load**

Cleaned, consistent Parquet files stored at silver/diabetes\_cleaned

**Creating a Single Note Cluster on Azure Data Bricks**

**A screenshot of a computer

AI-generated content may be incorrect.**

All-purpose cluster is created and started

**A screenshot of a facebook page

AI-generated content may be incorrect.**

## Bronze ➜ Silver

**Extract:**

**A screenshot of a computer

AI-generated content may be incorrect.**

Authenticating Azure Data Bricks with storage account through access-keys

Access-keys gives complete access to the storage account ( **puneethdiabetesstorage**)

**A screenshot of a computer

AI-generated content may be incorrect.**

Copied key1 from storage

**A screenshot of a computer

AI-generated content may be incorrect.**

Mounted storage to the Databricks mount using key1 copied . We can now access bronze files using “/mnt/diabetes/bronze” directly

Similarly we can access files from silver and gold layers using appropriate locations .

A screenshot of a computer

AI-generated content may be incorrect.

1. **Load Libraries and Define Schema**

To Define schema we imported functions from data frame Api of PySpark.  
  
Used Pyspark Coding instead of SQL to transform Raw data in the bronze Layer to cleaned and structured data

Defined schema of data by looking at top 5 rows of the dataset

**A screenshot of a computer

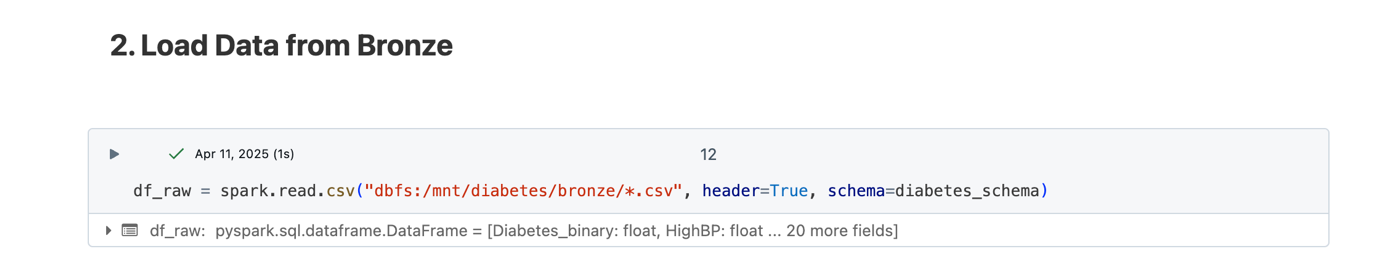
AI-generated content may be incorrect.**

A screenshot of a computer

AI-generated content may be incorrect.

1. **Load Data from Bronze**

Loaded the data from bronze layer into a data frame using spark read api and with schema : “diabetes\_schema” defined above.



1. **Inspect Raw Data**

A screenshot of a computer

AI-generated content may be incorrect.

Querying results using **display** and also inspecting data using **printSchema** and **describe** method

A screenshot of a computer

AI-generated content may be incorrect.

1. **Data Cleaning and Processing**

* Duplicate Values are dropped using **dropDuplicates()**

**A screenshot of a computer

AI-generated content may be incorrect.**

* BMI values like 98 are rare hence outliers are filtered.

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Save Cleaned Data to Silver Layer**

The raw data in the Bronze layer has now been structured using a well-defined schema. All missing and duplicate values have been addressed, ensuring clean and consistent data. To optimize storage and retrieval efficiency, the processed data is written to the Silver layer in Parquet format, a highly efficient columnar storage format.

A screen shot of a computer

AI-generated content may be incorrect.

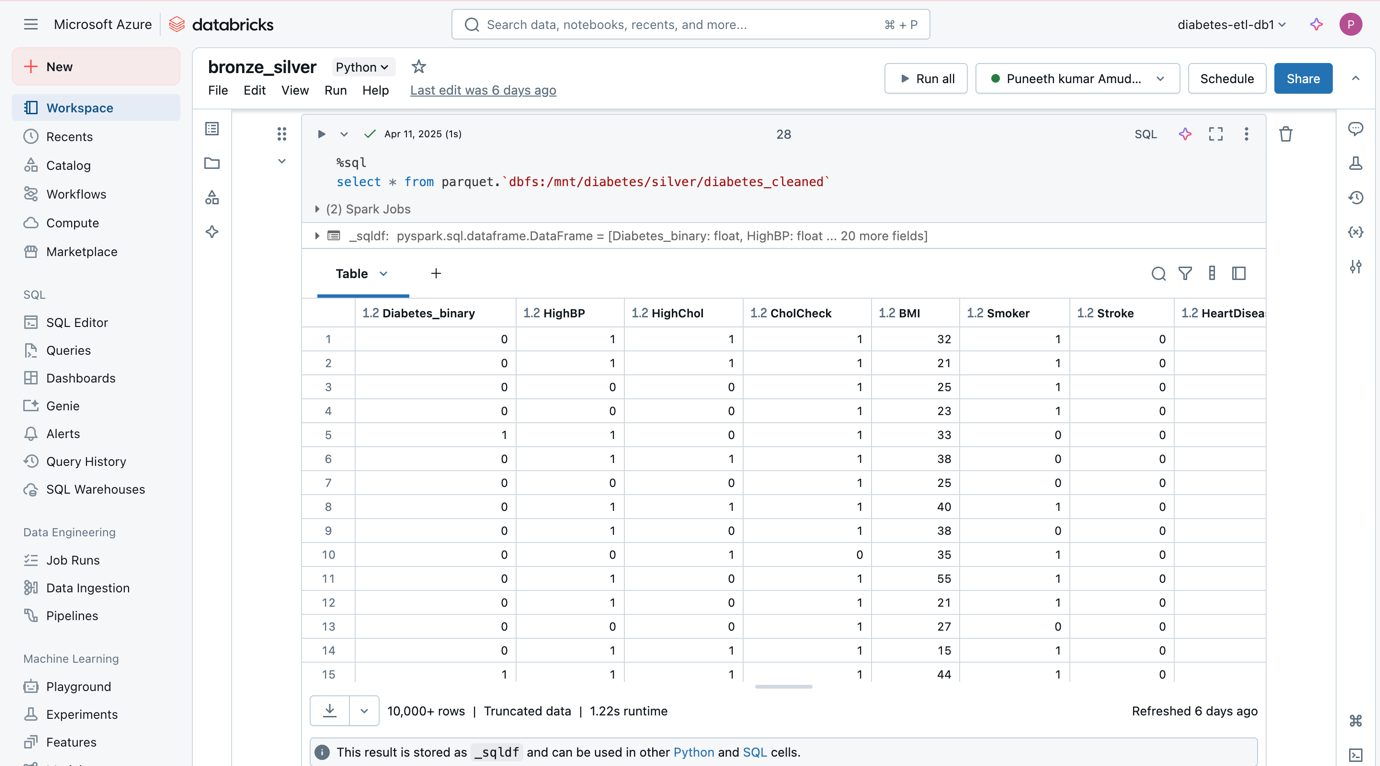
**Verification Queries -1(Bronze to Silver):**

**A screenshot of a computer

AI-generated content may be incorrect.**

The data is written to **Azure Data Lake Storage** under the storage account named “**puneethdiabetesstorage**”, specifically into the “**silver**” container. It is stored in Parquet format with partitioning applied to enhance query performance and scalability**.**

We can retrieve the data using **SQL queries** by referencing the **mount location** of the Azure Data Lake corresponding to the Silver layer.

****

**Phase-3 : Analytics Development**

## Silver ➜ Gold

Input: silver/diabetes\_cleaned

Key Steps:

* Feature engineering: "Obese", "HighMentalDistress", "HighPhysicalDistress"
* Transformation: Renaming columns
* Create a two algorithms : Risk Scoring and Population Segmentation
* Group by Age, Sex, and Risk Level to create summary metrics
* Store the aggregated results in Delta format

Output:

* Aggregated insights at gold/aggregated\_results

**Step-1:**

A screenshot of a computer

AI-generated content may be incorrect.

Named the notebook **“Silver\_to\_gold”** and loaded the data from the **silver container** using the **Databricks mount location**. Since the file is stored in **Parquet format**, specified the format as **“parquet”** while loading the data.

The original Age column has been scaled numerically from 1 to 12, where each number represents a 5-year age interval.

To improve interpretability and simplify analysis, grouped the scaled age values into four broader age groups

**A screenshot of a computer

AI-generated content may be incorrect.**

New columns called Obese, HighMentalDiStress and HighPhysicalDistress were created with certain assumptions

**A screenshot of a computer

AI-generated content may be incorrect.**

Created three new columns based on simple health rules:

• **Obese:** If BMI is 30 or more, mark as 1 (obese), else 0.

• **HighMentalDistress:** If MentHlth (mental health bad days) is more than 15, mark as 1, else 0.

• **HighPhysicalDistress:** If PhysHlth (physical health bad days) is more than 15, mark as 1, else 0.

These new columns help highlight people with potential health risks.

**A screenshot of a computer

AI-generated content may be incorrect.**

Renamed Column Called **HvyAlcoholConsump** to **AlcoholRisk**

**Step-2**

Creating Algorithms for further analysis:

**1. Develop Risk Scoring Algorithms**

**Goal** :

Assign a risk score to each individual based on lifestyle, mental/physical health, and chronic indicators

**Approach**: Weighted Composite Score

**A screenshot of a computer

AI-generated content may be incorrect.**

Found the correlation between each numeric column and the target column **Diabetes\_binary**. The correlation here is done to compute **weighted risk score.**A screenshot of a computer

AI-generated content may be incorrect.

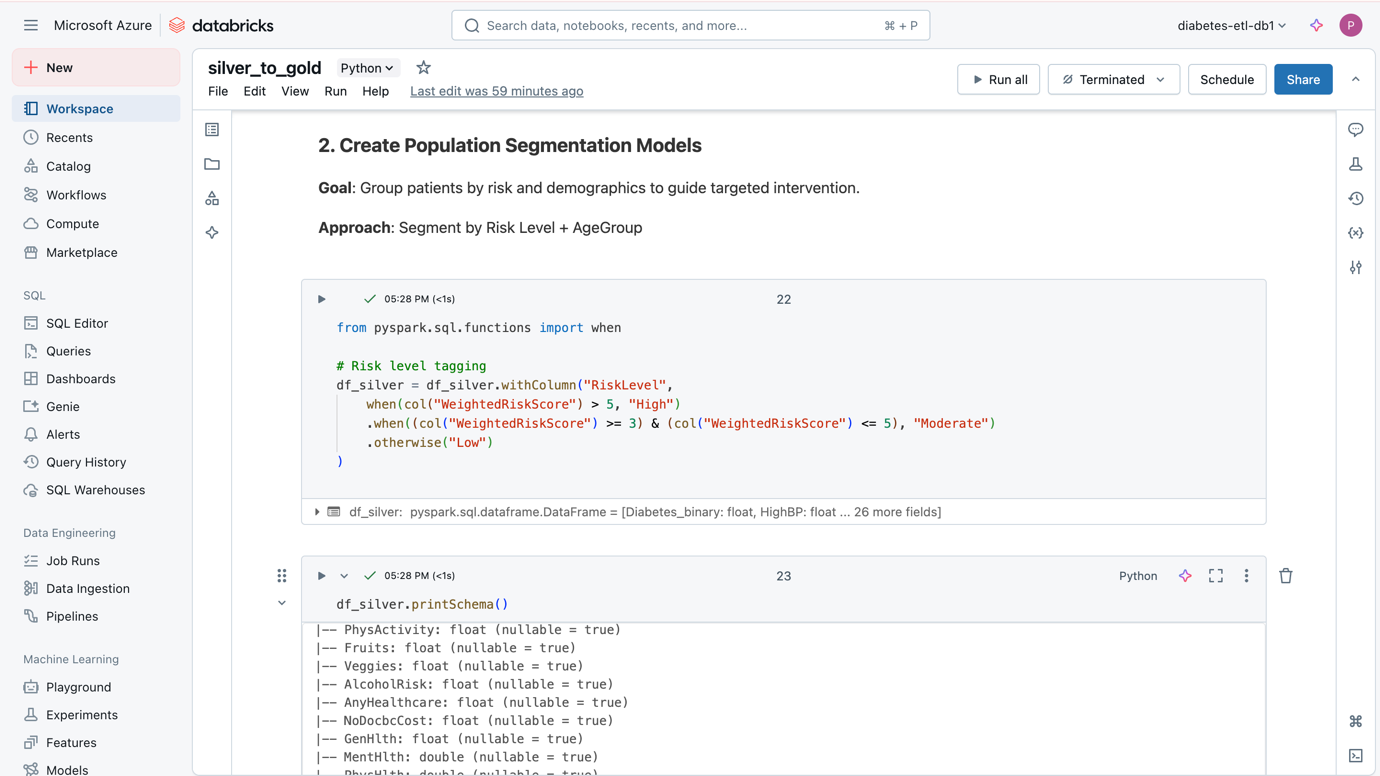
Assigned weight on basis of correlation and created a column called **“WeightedRiskScore”**

1. **Create Population Segmentation Models**

**Goal:**

Group patients by risk and demographics to guide targeted intervention.

**Approach:** Segment by Risk Level

****

Segmented patients by **WeightedRiskScore** into three **risk levels**: High, Moderate, and Low.

**Step-3:**

**Loading the data to Gold Container**In the Gold layer, we grouped patients by AgeGroup, Sex, and RiskLevel to get a summary of their health conditions. For each group, calculated the number of patients and the average rates of diabetes, obesity, low physical activity, alcohol risk, mental distress, and high blood pressure. These values help us understand health patterns in different groups and are useful for creating dashboards and reports.

A screenshot of a computer

AI-generated content may be incorrect.

New Data frame called **df\_gold** now contains all the aggregated data.  
  
Hence writing the data to the gold layer in **Delta** format with “**AgeGroup**” as Partition Column  
  
  
A screenshot of a computer

AI-generated content may be incorrect.

**Verification Query-2**

Data is now stored in gold container(layer) of Azure Data Lake Gen-2 storage called **puneethdiabetesstorage**.

A screenshot of a computer

AI-generated content may be incorrect.

The data is stored in **Delta format** and partitioned by the **AgeGroup** column. The presence of the **\_delta\_log** folder confirms that the dataset is managed as a Delta table, enabling features like versioning, efficient updates, and time travel.