Diabetic Data Analysis

By Carlos Velasco

Executive Summary

TA hospital system wants to analyze patient data to reduce readmission rates, optimize treatment outcomes, and detect patterns in hospital operations. The existing data is large and semi-structured, and the hospital wants to build a scalable data engineering pipeline using PySpark on Databricks:

- diabetic_data.csv: ~100,000 patient records.
- Data Details (Selected Columns):

  - race, gender, age
  - admission_type_id, discharge_disposition_id, admission_source_id
  - time_in_hospital, num_lab_procedures, num_procedures, num_medications
  - number_outpatient, number_emergency, number_inpatient • diag_1, diag_2, diag_3 (Diagnosis codes)
  - readmitted (Yes/<30/>30/No)
  - change, diabetesMed, insulin, A1Cresult

Objectives

1. Data Ingestion: :

• Load the CSV into Databricks using PySpark.

2. Data Cleansing:

• Handle missing values like "?" in race, gender, diag_1, etc.

• Remove or analyze invalid entries (e.g., gender = "Unknown/Invalid").

3. Data Transformation:

• Convert categorical columns into meaningful values (e.g., map admission types).

• Derive new columns such as:

o total_visits = number_outpatient + number_emergency + number_inpatient

o readmission_flag = 1 if readmitted in ("<30", ">30") else 0

4. Exploratory Analysis:

• Average time in hospital per age group or diagnosis.

• Identify which diagnoses have high readmission rates.

• Relationship between insulin/A1C levels and readmission.

Objectives

5. Performance Optimization:

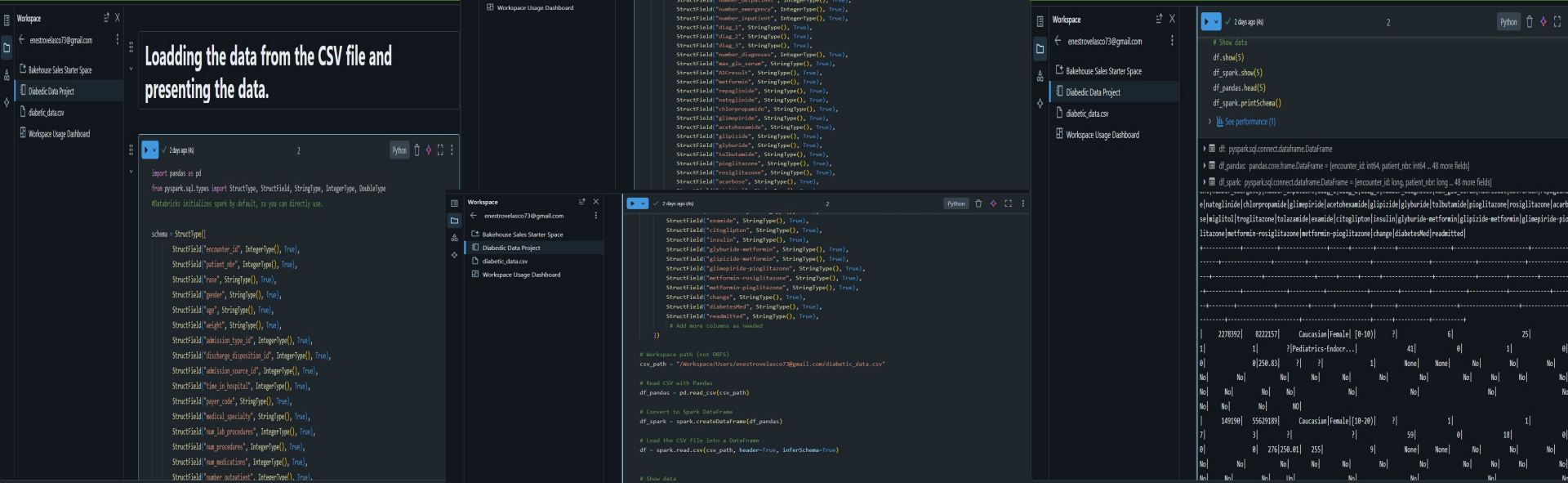
- Implement data partitioning by age or readmission_flag.
- Cache intermediate results for repeated analysis.

6. Data Export:

- Save cleaned dataset as Parquet/Delta.
- Export analytical summaries as CSV for BI tools.

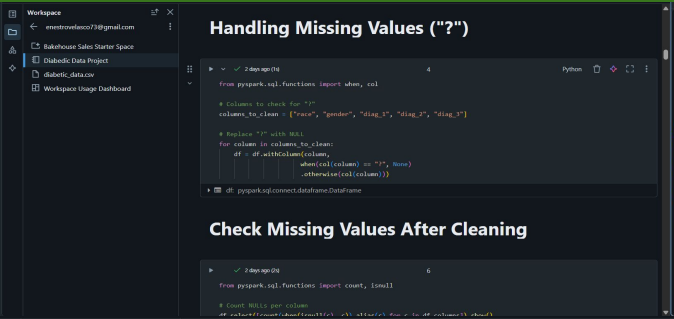
Data Ingestion

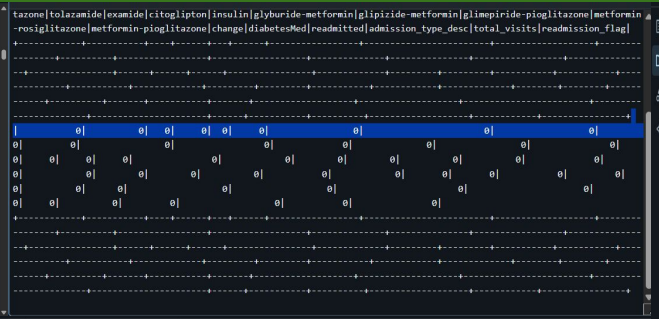
- Methodology: Load the CSV into Databricks using PySpark.
- Tools: PySpark
- Process: Data Collection
- Data Source:

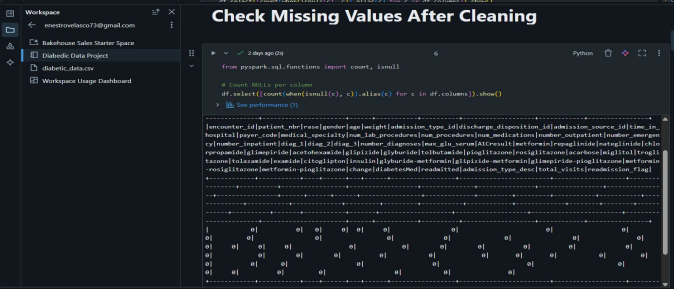


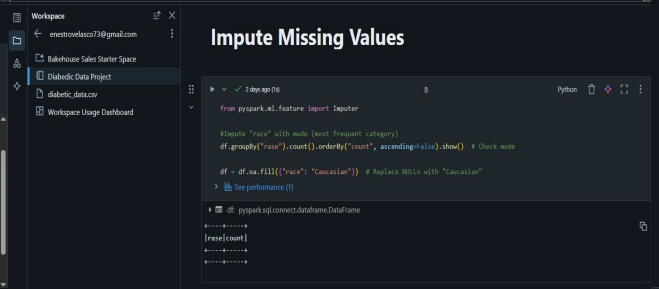
Data Cleaning

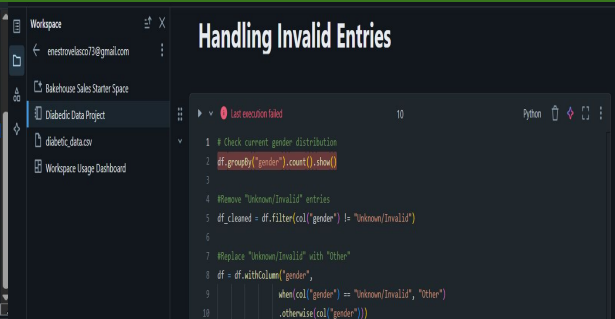
- Handle missing values like "?" in race, gender, diag_1, etc.
- Remove or analyze invalid entries (e.g., gender = "Unknown/Invalid").
- Data Source:









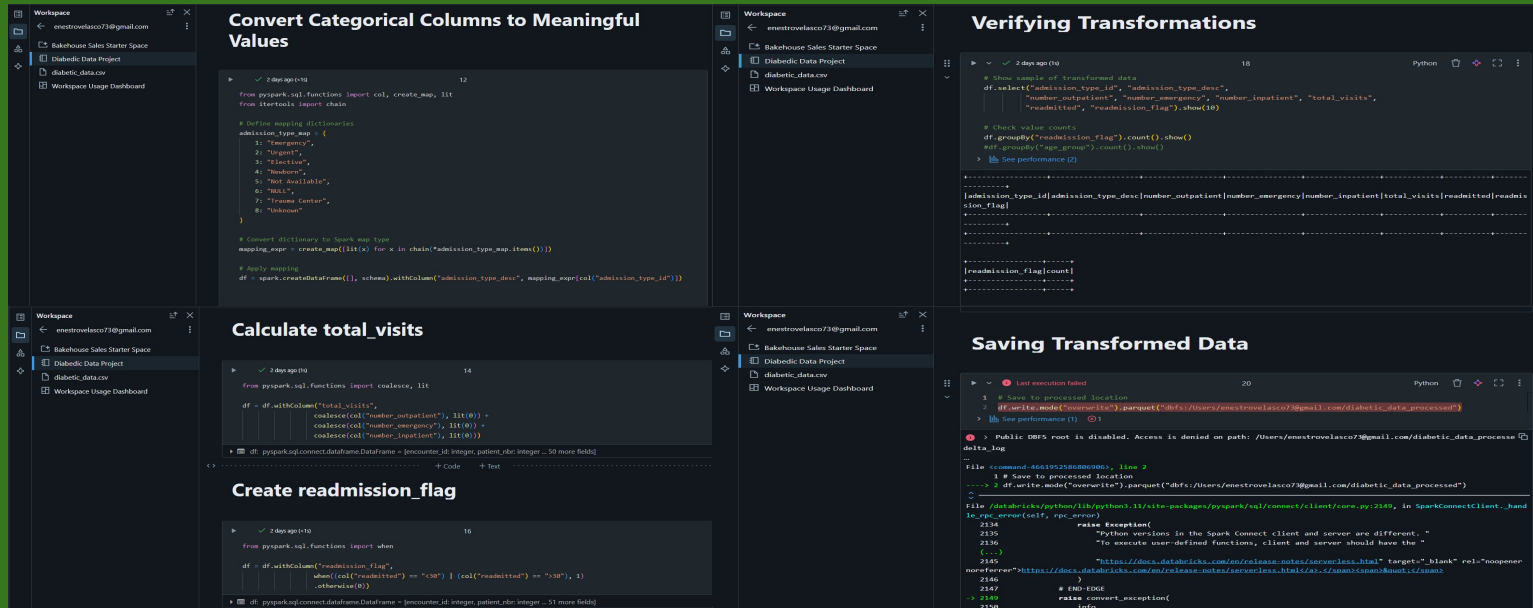


Data Transformation

- Convert categorical columns into meaningful values (e.g., map admission types).
- Derive new columns such as:

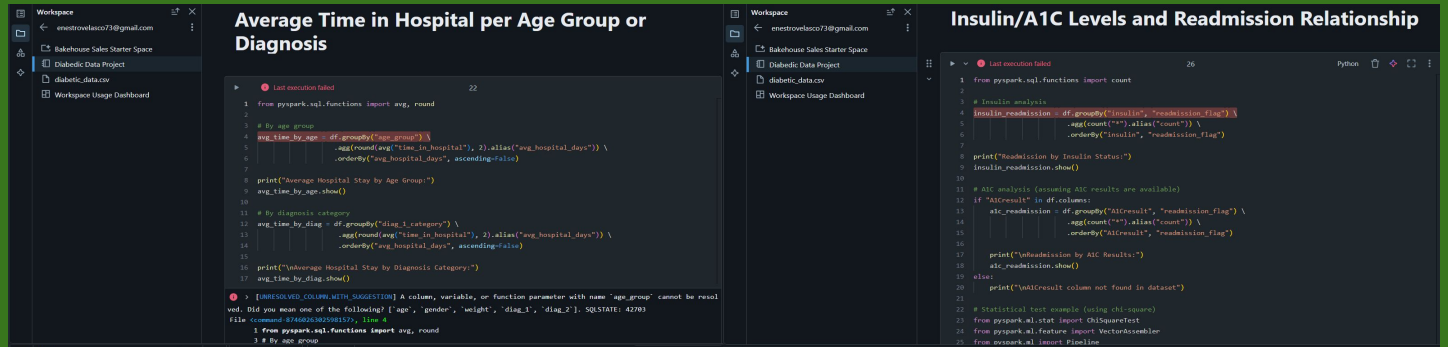
1) total_visits = number_outpatient + number_emergency + number_inpatient
2)readmission_flag = 1 if readmitted in ("<30", ">30") else 0

- Data Source:



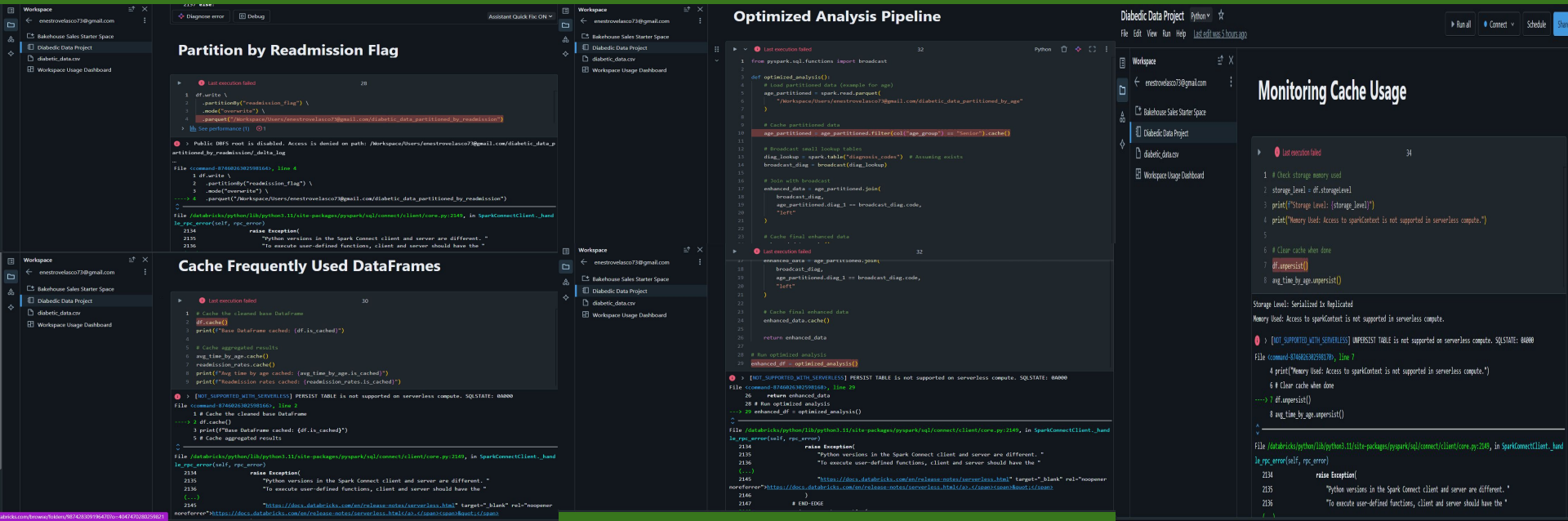
Exploratory Analysis

- Average time in hospital per age group or diagnosis.
- Identify which diagnoses have high readmission rates.
- Relationship between insulin/A1C levels and readmission.
- Data Source:



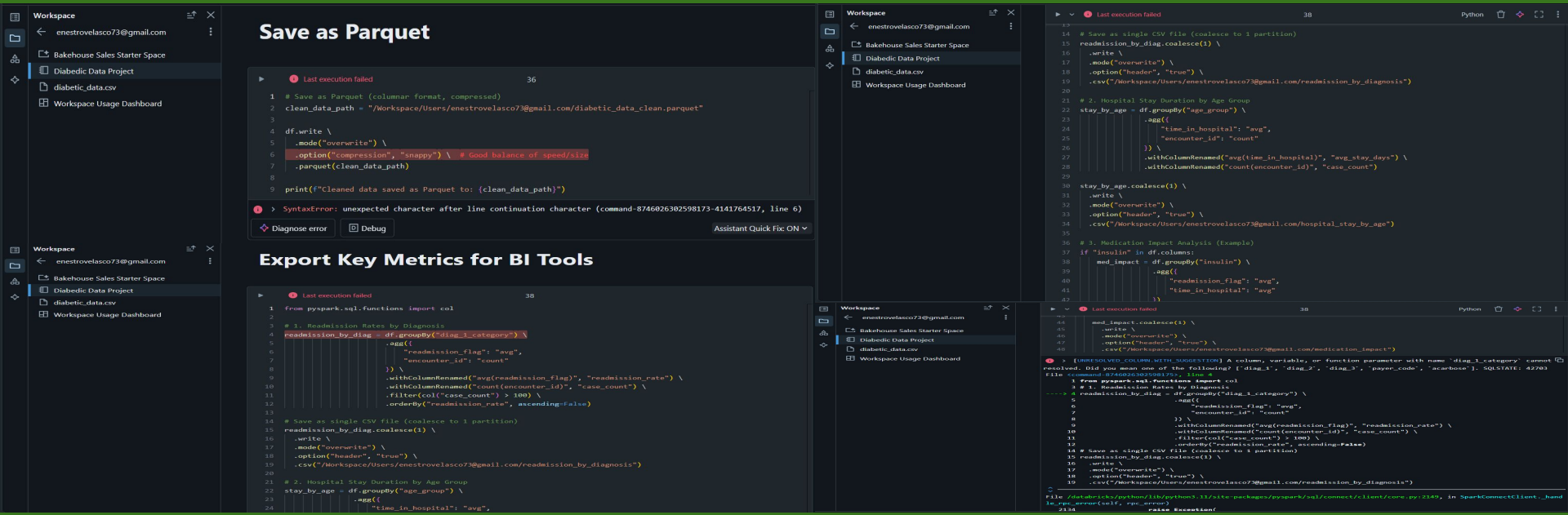
Performance Optimization

- Implement data partitioning by age or readmission_flag.
- Cache intermediate results for repeated analysis.
- Data Source:



Data Export

- Save cleaned dataset as Parquet/Delta.
- Export analytical summaries as CSV for BI tools.
- Data Source:



Conclusion

Completed Work: Was able to do the Data Ingestion, Data Cleansing, Data Transformation, Exploratory Analysis, Performance Optimization, Data Export. But, I did run into some connection errors in the code.

Moving forward: Will need to fix the DBFS root error I kept running into and the other errors I was running into.