NCPL

BOOTCAMP PROJECT 4

Submission by

Vinod Kumar Mannepuram

ETL Pipeline for Healthcare Data Using Azure Databricks

1. Project Overview

The ETL (Extract, Transform, Load) pipeline is designed to process healthcare data for a provider aiming to utilize large-scale patient records for predictive analytics and regulatory reporting. This project leverages a combination of Azure technologies and PySpark scripting within Azure Databricks environments.

2. Problem Statement

The healthcare provider seeks to leverage their vast repository of patient data to gain insights that would support decision-making and enhance patient care quality. The need to process and analyze data efficiently is crucial for their operational and strategic goals.

3. Technological Stack

* **Azure Databricks:** Provides the computational framework and environment for data processing using PySpark.
* **Azure Data Lake Storage (ADLS):** Used for the storage of raw data in a scalable and secure manner.
* **Azure Data Factory (ADF):** Orchestrates and automates data movement and transformation processes.
* **Azure SQL Database:** Serves as the data warehouse for storing processed data.
* **Python/PySpark:** Used for writing scripts to process and transform data.

4. Implementation Steps

**a. Data Ingestion**

Raw patient data is loaded from Azure Data Lake Storage using Azure Data Factory, which orchestrates the movement of data into Azure Databricks.

A screenshot of a computer

AI-generated content may be incorrect.

**b. Data Processing in Azure Databricks**

1. **Initial Data Loading:** Data is read from a CSV file stored in ADLS into a DataFrame.

python

Copy

df = spark.read.csv("/mnt/bootcamp4/Raw/Healthcare.csv", header=True)

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Data Transformation:** Includes cleaning and transforming data by:
   * Categorizing age groups.
   * Calculating costs per day of stay.
   * Encoding gender into numerical codes.
   * Creating aggregate metrics for each condition, such as average cost, average length of stay, and satisfaction.

A screenshot of a computer

AI-generated content may be incorrect.

1. **Data Aggregation:** Data is further processed to include summary statistics and prepared for loading into the Azure SQL Database.

python

Copy

df5 = df4.withColumnRenamed("Cost\_per\_Day", "Cost\_Per\_Day\_Adjusted")

1. **Data Loading:** The final DataFrame is written back to Azure SQL Database using JDBC connections secured by Azure Active Directory tokens.

python

Copy

df5.write.jdbc(url=url, table="dbo.Healthcare", mode="append", properties=properties)

**c. Data Output and Storage**

The transformed data is stored in Azure SQL Database, providing a reliable and secure platform for further analysis or reporting.

A screenshot of a computer

AI-generated content may be incorrect.

5. Validation and Quality Assurance

* **Data Quality Checks:** Implemented within the PySpark scripts to ensure the integrity of the data transformation.
* **Schema Validation:** Ensures that the data structures align with expected formats and types.

6. Deliverables

* PySpark scripts for data transformation.
* Transformed data stored securely in Azure SQL Database.
* Reports verifying data quality and transformation accuracy.

7. Project Completion Criteria

* Successful execution of data ingestion, transformation, and loading processes.
* Accurate and validated data available in the SQL database for analytics and reporting.
* Enhanced capabilities for the healthcare provider to perform data-driven decision-making.

This document encapsulates the workflow and methodologies used in the ETL pipeline development for handling healthcare data, highlighting the integration of Azure services and Databricks for robust data processing and analysis.