



| University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Master of Science in Data Science

TECHNICAL REPORT



SEMESTER - 2

CONTENTS

1

Project Name.....	2
Executive Summary	2
Technical Report.....	3
Highlights of Project	4
Submitted on:.....	4
Abstract.....	5
Methodology.....	7
Data Flow Description.....	8
Results Section	9
Discussion.....	11
Conclusion	12
Contributions/References.....	13

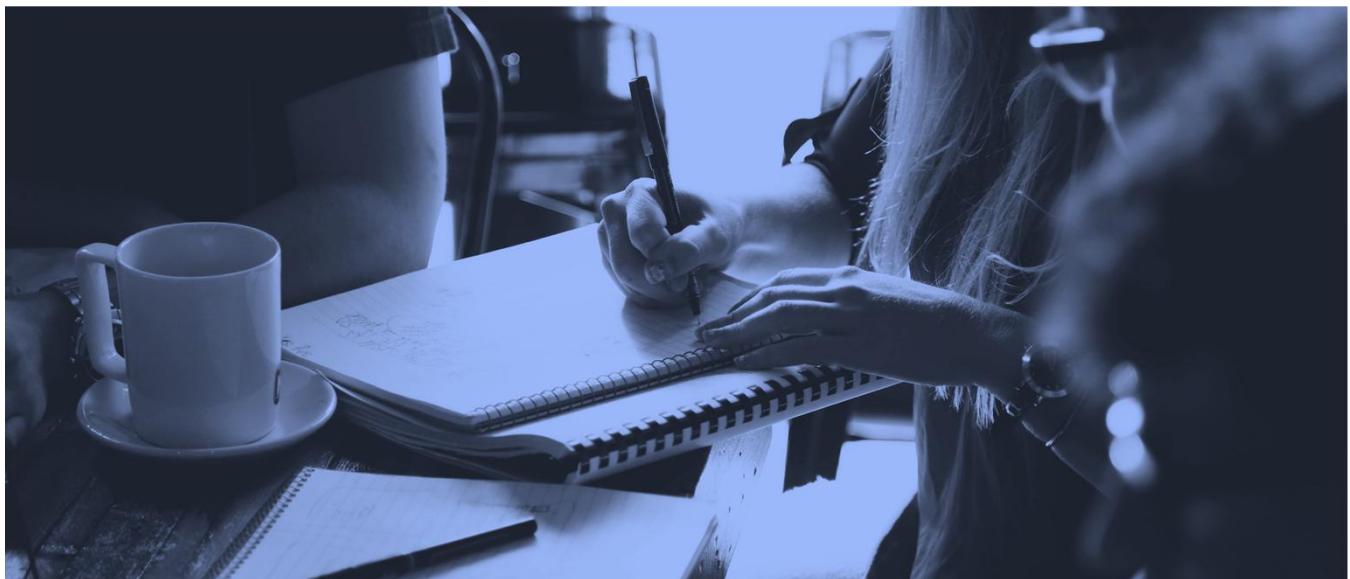
Project Name

Executive Summary

This project focuses on helping students better understand healthcare costs by creating a clear, automated, and easy-to-use data system. Many students delay or avoid medical care because costs and insurance details are confusing. To solve this problem, our team designed a fully automated data pipeline using AWS services such as S3, Lambda, Glue, Crawler, and Athena. The system collects raw healthcare claim data, cleans it, transforms it, and stores it in a centralized location.

Students can explore the final results through an interactive Power BI dashboard, where they can view important insights like total paid amounts, billed amounts, top diagnoses, and provider types. The system is fast, scalable, and requires little manual work. It ensures transparency in healthcare costs and supports students in making informed health decisions.

This report explains the problem we addressed, the methods we used, the results produced, and the lessons we learned while completing this project.



Technical Report

Team Members:

Kavya Bommineni

Deepak Avinash Katuri

Trinath Guru

Satish Kumar Jonna

Questions?

Contact : dkatu1@unh.newhaven.edu

Promoting Transparent and Predictable Healthcare Costs for Students.

Highlights of Project



- Our project solves the problem of unclear healthcare costs for students by making pricing information simple and predictable.
- We built an automated AWS pipeline where uploading a CSV triggers Lambda and starts a Glue ETL job.
- Glue cleans, transforms, and stores the data in Parquet format for fast querying.
- A Glue Crawler updates the Catalog so Athena can instantly query the latest data.
- Power BI displays easy-to-read visuals like paid amounts, member counts, diagnoses, and provider types.
- The system requires almost no manual work and can scale as more data is added.
- We learned about data quality, healthcare terminology, AWS integration, and building end-to-end automated pipelines.

Abstract

This project provides a scalable and automated solution to improve healthcare cost transparency for students. Many students avoid medical treatment due to not knowing how much they may need to pay and not understanding insurance terms. To address this issue, we built an automated data engineering pipeline using AWS services, including S3, Lambda, Glue, Crawler, and Athena.

The pipeline processes raw healthcare claim data into clean, structured information that can be easily analyzed. An interactive Power BI dashboard displays key insights such as paid amounts, billed amounts, provider types, diagnosis trends, and year-wise member counts. The system runs automatically with minimal manual work and ensures accuracy through proper data cleaning and transformation.

Overall, this solution gives students simple, clear access to healthcare cost information, helping them make informed decisions. The pipeline is scalable, reliable, and ready for future expansion such as real-time analytics and predictive modeling.

Introductory Section

Students often struggle to understand healthcare costs because medical pricing and insurance coverage are not explained clearly. This lack of transparency creates confusion and leads many students to delay necessary medical care. To help solve this problem, our project focuses on making healthcare cost information easy to access and simple to understand. We built an automated system that collects, cleans, and presents healthcare data in a clear way so students can make informed decisions without stress.

Students today often come from different countries and may not be familiar with how the U.S. healthcare system works, which makes the confusion even greater. Without clear cost information, they cannot plan their budgets or understand what services they can afford. Our project aims to bridge this gap by providing a clear, automated, and easy-to-use system that helps students see their healthcare costs in a straightforward way.

Review of available research

Research shows that lack of price transparency in healthcare is a major barrier for patients, especially young adults and students. Many studies highlight how unclear medical costs contribute to delayed treatment, financial anxiety, and poor decision-making. Existing tools often provide high-level cost estimates but lack personalized or easy-to-read information. This gap shows a need for systems that simplify healthcare cost data and make it understandable for non-experts. Our project builds on this idea by creating a transparent, automated data pipeline and dashboard that helps students view important cost information in a clear and accessible way.

Methodology

We followed a structured approach to build our solution, focusing on automation, clarity, and user accessibility.

CRISP-DM Methodology

Business Understanding:

Students face confusion and uncertainty about healthcare costs. There is no simple way for them to view or understand out-of-pocket expenses, which leads to delays in seeking care.

Data Understanding:

We used synthetic healthcare claim data representing member IDs, provider types, provider specialties, diagnoses, paid amounts, billed amounts, and dates. We analyzed the structure, missing values, and inconsistencies.

Data Preparation:

Data was cleaned using AWS Glue by removing unnecessary fields, handling missing values, renaming columns, adding timestamps, and converting CSV data into Parquet format for efficient querying.

Modeling:

Although we did not build predictive models, we created data transformations and SQL queries in Athena to support strong analysis and visual insights.

Evaluation:

We validated the pipeline by testing multiple uploads, checking data accuracy in Glue and Athena, and matching dashboard values with expected totals

Data Flow Description

Step 1: Raw Data Ingestion

- Raw healthcare claims data (CSV format) is uploaded manually to an Amazon S3 bucket under the raw/ prefix.
- The CSV contains fields such as MEMBER_ID, PROVIDER_NPI, CLAIM_ID, DIAGNOSIS_CODES, SERVICE_DATES, ALLOWED_AMOUNT, PAID_AMOUNT, and other medical billing details.

Step 2: Event Trigger

- Amazon Lambda Function is configured to detect the new object upload event.
- Upon file upload, the Lambda Function triggers an AWS Glue ETL job.

Step 3: AWS Glue ETL Job - Data Cleansing and Transformation

- The Glue job reads raw CSV files from the raw S3 path.
- Cleansing operations performed:
 - Dropping unwanted columns.
 - Renaming columns name.
 - Adding a column with current date.
 - Remove incomplete records (null or invalid values).
 - Standardize date formats (e.g., yyyy-mm-dd).
 - Ensure numeric consistency in financial fields (e.g., ALLOWED_AMOUNT, PAID_AMOUNT).
 - Validate critical identifiers like PROVIDER_NPI (should be 10 digits).
 - Standardize column names and fix schema inconsistencies.
- The transformed data is written back to S3 under a separate cleaned/ prefix bucket.

Step 4: AWS Glue Crawler

- A scheduled AWS Glue Crawler scans the cleaned/ S3 path.
- The crawler updates the AWS Glue Data Catalog with the latest schema of the cleaned data.

Step 5: Amazon Athena

- Amazon Athena is configured to query the cleaned data directly from S3.
- SQL queries can be performed for analytical purposes such as claim aggregation, provider payment summaries, or diagnosis-based insights.

Step 6: Power BI Dashboard

- Power BI connects to Athena using the ODBC or JDBC connector.
- Dashboards visualize key KPIs like:
 - Total Claims Paid o Top Service Providers
 - Claims Volume by Diagnosis Code
- Power BI auto-refreshes dashboards when new clean data is added.

Results Section

Overview

The automated pipeline successfully processed all uploaded CSV files and converted them into clean Parquet datasets. Athena was able to run fast SQL queries on the transformed data, showing accurate results.

Key Findings

- Total paid amount and total billed amount were clearly displayed using card visuals.
- Provider type and provider specialty distributions helped identify where most claims originated.

- Diagnosis trends showed which health issues were most common among students.
- Year-wise member counts helped analyze healthcare usage over time.

Data Engineering Pipeline

Data Ingestion:

- Raw data uploaded into S3 source bucket
- Triggered automatically through Lambda

Data Storage:

- Storage in S3 using two buckets (raw + transformed)
- Use of Parquet format for optimized storage

Data Processing:

- AWS Glue ETL script performed cleaning, transformations, field removal, renaming, and timestamp creation

Data Consumption:

- Athena enabled explore-and-query
- Power BI created easy-to-understand visual insights

Model Deployment:

Not applicable, but system prepared for future model integration.

Data Visualization:

Power BI dashboard displayed all insights, including bar charts, pie charts, donut charts, year-wise trends, and total financial amounts.

Workflow of Project

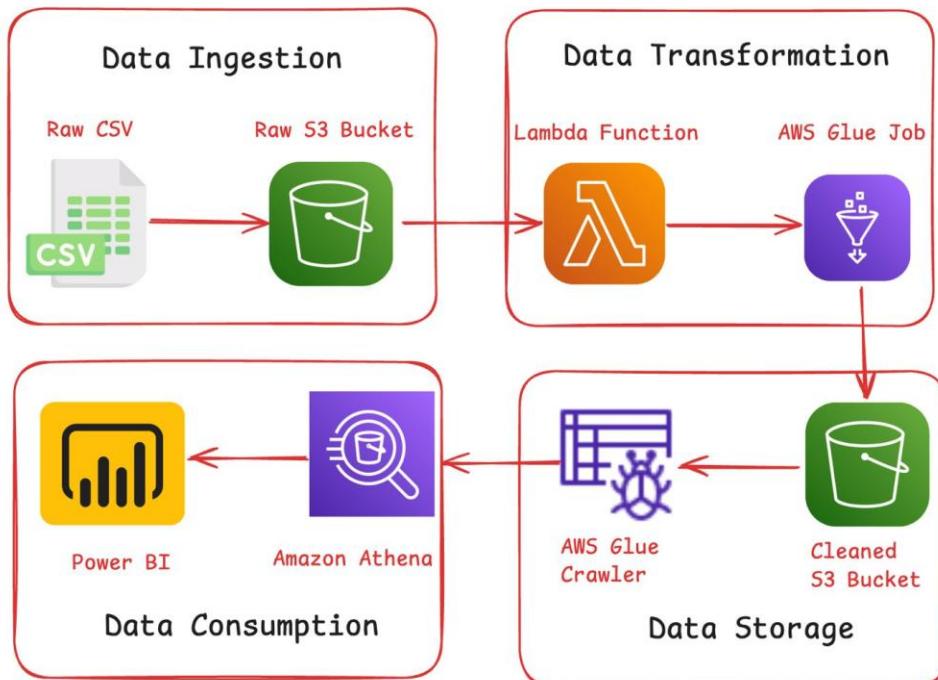


Fig-1

Discussion

This project set out to tackle the challenge faced by international students in understanding healthcare expenses in the U.S.—a system often marked by complexity and lack of transparency. By building an automated data pipeline and Power BI dashboards, the project provides a practical solution that turns raw claims data into clear, actionable insights.

The results show that students can now view summaries of total claims, top providers, and common diagnoses, helping them make better decisions around

healthcare usage and cost. This addresses a key knowledge gap, making healthcare billing more understandable and accessible.

However, while the solution simplifies the data, it doesn't fully demystify insurance terms, coverage rules, or out-of-pocket costs, which remain complex and outside the scope of this pipeline. Additionally, scaling to more diverse or real-time data sources may require further development.

Still, this project offers a strong foundation for improving healthcare cost transparency, especially for underserved users like international students. It's a step toward empowering users with data-driven tools to navigate a confusing system more confidently.

Conclusion

This project successfully delivers an end-to-end, automated data pipeline that transforms raw healthcare claims data into meaningful insights—empowering international students to better understand their medical expenses. By leveraging AWS services like S3, Glue, and Athena, and visualizing the results through Power BI, the solution provides a scalable and efficient way to increase healthcare cost transparency.

While it may not cover every complexity of the U.S. healthcare system, it represents a significant step toward improving digital health literacy. The system not only reduces manual effort but also offers a platform for future enhancements, including machine learning and deeper policy integration.

Ultimately, this work lays a strong foundation for leveraging data engineering and analytics to bridge information gaps and empower informed healthcare decisions, especially within vulnerable communities such as international students.

Contributions/References

KAVYA BOMMINENI

DEEPAK AVINASH KATURI

TRINATH GURU

SATISH KUMAR JONNA