|  |
| --- |
| ANALYZING STUDENTS PERFORMANCE DATA |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| Fall 2024 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

|  |
| --- |
| Analyzing Students Performance Data |

|  |  |  |
| --- | --- | --- |
| Executive Summary Developed a scalable cloud-based pipeline using AWS to analyze student performance data, providing actionable insights to enhance educational outcomes. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Ashok Kumar Jugurubilli**  **Kavya Rampalli**  **Suman Bhattarai** | **Questions?**  Contact :  Ajaru1[@unh.newhaven.edu](mailto:nadla1@unh.newhaven.edu)  Kramp2[@unh.newhaven.edu](mailto:evega2@unh.newhaven.edu)  [sbhat19@unh.newhaven.edu](mailto:sbhat19@unh.newhaven.edu) |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| **Analyzing Students Performance Data** |  |
| Submitted on: 12/8/2024 |

## Abstract

**Abstract**

This project focuses on leveraging cloud-based tools and data analytics techniques to analyze student performance data and identify trends, insights, and opportunities for improvement in educational outcomes. The workflow is built around AWS services for scalable and efficient data handling, ensuring streamlined data ingestion, transformation, querying, and visualization.

The process begins with uploading raw student data into Amazon S3, a secure and scalable storage service. Data preprocessing is performed using AWS Lambda, a serverless computing platform, which allows for efficient transformation of raw data into a clean, structured format. The transformed data is then stored in a separate S3 bucket to maintain data segregation and facilitate efficient querying.

AWS Athena, a serverless interactive query service, is employed to execute SQL queries directly on the stored data in S3. This enables flexible and on-demand analysis of student data without the need for complex infrastructure management. The query results are exported as CSV files for further analysis.

Using Python in Jupyter Notebook, the project visualizes the processed data to highlight key performance metrics, trends, and correlations. Graphical representations, such as line plots, scatter plots, box plots and histograms, offer insights into factors impacting student performance, such as attendance, subject-wise scores, and demographic patterns.

The architecture is designed with scalability and cost-effectiveness in mind, incorporating best practices such as data partitioning and the use of optimized data formats (e.g., Parquet or ORC) to enhance query performance. Security measures, including IAM roles, data encryption, and access controls, ensure the safe handling of sensitive student data.

This project serves as a foundation for advanced analytics and machine learning, enabling predictive modeling for identifying at-risk students, tailoring interventions, and improving overall educational outcomes. By automating the data pipeline and integrating modern visualization tools, this approach provides educators and stakeholders with actionable insights to foster data-driven decision-making in education.

Add your elevator pitch video link here ()

Introductory Section

The evaluation of student performance is a critical process for educational institutions, as it informs decision-making, curriculum development, and targeted interventions. However, traditional methods of collecting, analyzing, and interpreting performance data can be inefficient and prone to human error. In many cases, data is siloed across disparate sources, lacks standard formatting, and is not readily accessible for in-depth analysis. This fragmented approach hinders the ability of educators and administrators to gain a comprehensive understanding of student achievement and identify areas that require improvement.

To address these challenges, there is a need for a scalable, efficient, and secure solution that leverages modern data analytics and cloud computing technologies. Such a solution can streamline the entire data lifecycle—from raw data ingestion and preprocessing to flexible querying and interactive visualization. By harnessing these capabilities, educators and decision-makers are better equipped to uncover meaningful insights, track performance trends over time, and ultimately enhance the quality of education.

In this technical report, we present a data pipeline and analysis framework built around Amazon Web Services (AWS) that provides a robust, reliable, and cost-effective means of managing student performance data. This solution leverages key AWS services including Amazon S3 for data storage, AWS Lambda for serverless preprocessing, AWS Athena for interactive querying, and Jupyter Notebook-based visualization tools for presenting insights. By automating data workflows, applying standardized transformations, and employing serverless technologies, this approach ensures that data is always accessible, accurate, and ready for analysis.

Through the following sections, we detail the methods used to implement the pipeline, including the architectural decisions, data handling strategies, and analytical techniques. We also discuss the results, insights gained, and recommendations for future enhancements. The overarching goal is to empower educational stakeholders with actionable intelligence, enabling data-driven decision-making that supports continuous improvement in student outcomes.

## Methodology

## Business Understanding

## The goal of the project is to improve e-commerce customer targeting and retention strategies by segmenting customers based on their behavior and preferences. This segmentation will help the business achieve:

## Enhanced personalized marketing efforts.

## Improved resource allocation for marketing campaigns.

## A better understanding of customer lifetime value (CLV) and retention drivers.

## Key Research Questions:

## How can customer behavior patterns be identified from transactional data?

## What are the key features that distinguish customer segments in the dataset?

## How can actionable insights be derived to optimize marketing strategies?

## Bolstered by Literature Review: Research in customer segmentation highlights the importance of RFM (Recency, Frequency, Monetary) analysis and clustering algorithms (e.g., k-means, hierarchical clustering) as effective methods to identify customer groups.

1. **Data Ingestion and Storage:**The initial step involved collecting raw student performance data—such as examination scores, attendance records, and demographic information—and uploading it to an Amazon S3 bucket. S3 was chosen for its scalability, durability, and integration capabilities with other AWS services. This facilitated the efficient ingestion of large volumes of data from various sources, while ensuring that the data remained secure and highly available.
2. Data Preprocessing and Transformation:  
   Next, the raw data underwent a preprocessing phase to clean, standardize, and structure it for downstream analysis. AWS Lambda, a serverless computing service, was employed to orchestrate these data transformations. The Lambda functions performed tasks such as handling missing values, converting formats (e.g., CSV to a columnar format), and removing duplicate records. By running transformations on-demand, Lambda eliminated the need for dedicated server infrastructure and allowed for cost-effective scaling as the volume of data evolved.
3. Structured Data Storage:  
   After preprocessing, the transformed data was stored in a dedicated S3 bucket. This separation of raw and cleaned datasets was intentional, providing clear data lineage and ensuring that the analysis could be repeated or audited at any point. It also simplified data management and version control, allowing both historical and updated datasets to coexist without confusion.
4. Querying and Analysis Using AWS Athena:  
   To analyze the structured data, AWS Athena was employed to run SQL queries directly against the dataset residing in S3. Athena’s serverless model allowed queries to scale automatically without the need to provision or manage servers. By defining schemas and tables through the AWS Glue Data Catalog, the dataset was made easily queryable. This enabled both exploratory analysis and more complex joins or aggregations to identify trends in student performance, correlate factors like attendance with grades, and pinpoint areas in need of intervention.
5. Exporting Results for Further Analysis:  
   After refining the queries and extracting meaningful insights, the resulting data subsets were saved as CSV files. These CSV files were then downloaded for further processing and visualization. By maintaining a clear workflow from raw data to analysis-ready outputs, it became straightforward to iterate on analytical steps, refine queries, or introduce new variables into the investigation.
6. Visualization and Reporting in Jupyter Notebook:  
   The downloaded CSV files were loaded into a Jupyter Notebook environment for visualization using Python data analysis libraries (e.g., pandas, matplotlib). Visualizations such as bar charts, scatter plots, and heatmaps helped transform raw metrics and query results into intuitive graphs, enabling easy interpretation by stakeholders. Through iterative analysis, different metrics could be tested and compared, facilitating data-driven discussions about student performance.
7. Validation and Continuous Improvement:  
   Throughout the methodology, iterative validation was conducted to ensure data integrity and analytical accuracy. Insights gleaned from the initial analysis were used to refine preprocessing steps, adjust queries, and experiment with alternative visualization techniques. Feedback loops between educators, administrators, and analysts helped hone the approach, ensuring that findings remained relevant, accurate, and actionable.

This methodology ensured an end-to-end pipeline for working with student performance data that was both technically sound and strategically aligned with the goal of improving educational outcomes. By leveraging AWS’s serverless and scalable offerings, the workflow minimized maintenance overhead, reduced costs, and delivered timely, meaningful insights to decision-makers.

## Results Section

Below is an outline of the methodology for the project, framed within the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework:

Business Understanding

* The primary goal of this project is to leverage student performance data to inform data-driven decision-making in educational settings. Educational institutions seek insights into factors influencing student outcomes, including subject-wise proficiency, attendance patterns, demographic influences, and the effectiveness of interventions. By understanding these elements, stakeholders can identify areas for academic support, curriculum enhancement, and strategic resource allocation. The ultimate aim is to improve overall student achievement, bolster retention rates, and guide effective policy development.

Data Understanding

The data used in this project typically comes from multiple sources, including student information systems, learning management platforms, and standardized assessment records. Initially, the raw datasets may exhibit inconsistent formats, missing values, or duplicated entries. To gain a comprehensive understanding, the following steps are undertaken:

* **Data Exploration:** Examining raw records, attributes, and their distributions.
* **Quality Checks:** Identifying anomalies, missing values, and potential outliers.
* **Contextual Insight:** Understanding the meaning of each variable, relevant time frames, and how these data points relate to real-world educational processes.
* **Data Preparation**
* Data preparation involves transforming the raw, heterogeneous data into a clean, consistent, and analysis-ready format. This stage includes:
* **Data Cleaning:** Handling missing values, correcting inaccuracies, and removing duplicates.
* **Feature Engineering:** Creating new variables or extracting meaningful features (e.g., attendance rates, average subject scores, or performance trends over time).
* **Data Integration:** Merging multiple datasets into a single, cohesive dataset.
* **Data Formatting:** Converting the data into suitable storage formats and partitioning it to improve query performance and scalability.

Modeling

The modeling phase focuses on employing analytical and statistical techniques—potentially including machine learning models—to derive actionable insights. Depending on the project’s objectives, this may involve:

* **Predictive Modeling:** Using regression, classification, or tree-based models to predict student performance or identify at-risk students.
* **Descriptive Analytics:** Employing clustering, factor analysis, or correlation studies to detect patterns and relationships.
* **Prescriptive Insights:** Recommending interventions or resources based on model outcomes (e.g., suggesting remedial classes for low-performing groups).
* **Evaluation**
* Once models are developed, their performance and the overall value they provide must be critically assessed. Evaluation includes:
* **Model Performance Metrics:** Analyzing accuracy, precision, recall, F1 scores, or other relevant metrics if predictive modeling is involved.
* **Statistical Significance:** Ensuring identified patterns are statistically meaningful rather than coincidental.
* **Interpretability & Practical Utility:** Confirming that the insights are understandable, actionable, and align with the institution’s objectives.
* **Continuous Improvement:** Identifying ways to refine data collection, preparation, and modeling methods in subsequent iterations.
* By following the CRISP-DM methodology, the project ensures a structured, iterative approach to data analysis—transforming raw student performance records into valuable insights that support informed decision-making and drive continuous improvement in the educational environment.

## Discussion

The project demonstrates the power of a cloud-based pipeline for analyzing student performance data, integrating various AWS services to streamline data handling, processing, and analysis. By leveraging Amazon S3 for data storage, AWS Lambda for preprocessing, and AWS Athena for querying, the workflow efficiently handles raw, unstructured data and transforms it into actionable insights. This approach eliminates the need for manual intervention in data cleaning and processing, significantly reducing errors and improving overall efficiency.

One of the most significant outcomes of this project was the ability to identify patterns and trends in student performance through structured analysis. For example, SQL queries on the transformed data provided insights into subject-wise performance, attendance trends, and demographic influences. These results were further enhanced through Python-based visualization tools in Jupyter Notebook, enabling stakeholders to interpret complex data with ease.

Challenges during the project included handling missing values, standardizing inconsistent data formats, and ensuring that the data pipeline could scale for larger datasets. These challenges were addressed through careful preprocessing using AWS Lambda and optimizing query performance in Athena by leveraging partitioning and efficient data formats like Parquet.

While the pipeline was successful in uncovering key insights, there is room for further enhancement. Integrating machine learning models through AWS SageMaker could enable predictive analytics, such as identifying at-risk students or recommending personalized interventions. Additionally, the inclusion of interactive dashboards using tools like AWS QuickSight could make insights more accessible to educators and administrators.

In conclusion, this project highlights the importance of leveraging modern cloud-based tools to handle complex datasets in an educational context. By automating data preparation and analysis, the pipeline not only saves time and resources but also empowers institutions with the insights needed to improve student outcomes and educational strategies. Future work can expand the scope of this framework to incorporate predictive and prescriptive analytics, further advancing its utility for data-driven decision-making in education.

## Conclusion

The project successfully demonstrates the use of a cloud-based data pipeline to analyze student performance data, uncovering valuable insights and trends that support data-driven decision-making in education. By leveraging AWS services such as S3, Lambda, and Athena, the workflow efficiently handled data ingestion, preprocessing, and querying, transforming raw, inconsistent data into structured, analysis-ready datasets. The visualizations created using Python in Jupyter Notebook provided clear and actionable insights for stakeholders, enabling targeted interventions and strategic improvements.

This approach not only improved the efficiency of handling large datasets but also laid the foundation for advanced analytics and predictive modeling. The project highlights the potential for integrating machine learning tools and interactive dashboards to further enhance insight generation and accessibility. Overall, the solution empowers educational institutions to optimize strategies, address learning gaps, and foster continuous improvement in student outcomes.