A black background with white text

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| Module Title | Big Data And Cloud Computing |
| Module code | LDS7005M |
| Student ID | 230264357 |
| Submission Date | 4 June 2024 |

BIG DATA & CLOUD COMPUTING

**Table of Contents**

[Introduction 3](#_Toc168332693)

[Data Ingestion and Storage 3](#_Toc168332694)

[Data description 3](#_Toc168332695)

[Proposed cloud based solution 4](#_Toc168332696)

[Data ingestion process 5](#_Toc168332697)

[Scalable Processing Architecture 6](#_Toc168332698)

[Big data processing architecture 6](#_Toc168332699)

[Choice of cloud services 6](#_Toc168332700)

[Data Extraction and Pre-processing 7](#_Toc168332701)

[Data extraction process 7](#_Toc168332702)

[Data Analysis and Insights 9](#_Toc168332703)

[Cloud-based analytics solution for analyzing 9](#_Toc168332704)

[Machine learning algorithms applied 10](#_Toc168332705)

[Analysis contribution to informed decision-making 10](#_Toc168332706)

[Cost Optimization Strategies 11](#_Toc168332707)

[Strategies for optimizing costs associated 11](#_Toc168332708)

[Security and Compliance 11](#_Toc168332709)

[Security measures and compliance considerations 11](#_Toc168332710)

[Performance Monitoring and Management 12](#_Toc168332711)

[Outline strategies for monitoring the performance of big data 12](#_Toc168332712)

[Exploring tools 12](#_Toc168332713)

[References 13](#_Toc168332714)

# Introduction

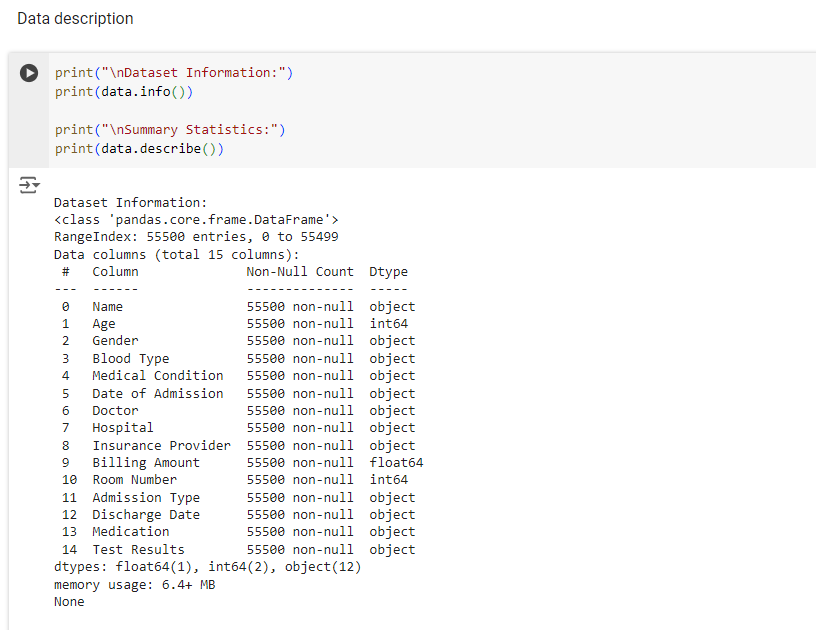
In this study of entire analysis, the data analysis and visualization have a strong relevance to the concept of value in the healthcare domain by helping to increase patients’ quality of life and organizational effectiveness. In data preprocessing, the first step includes renaming all the columns as per the standard naming convention and the missing values also handled at this stage Categorical data has been encoded using features like stay duration and new features of interest have also been created . A model using Random Forest classifier was fit to train the classifier to predict medical conditions of a patient based on various attributes of the patient. The outcomes captured in the form of the accuracy measures reveal the efficacy of the model suggesting its applicability for improving prognostic assessments in healthcare delivery systems.

# Data Ingestion and Storage

## Data description

The description of the entire dataset has been discussed in this section where the entire process has been described. However, the initial stage and vital stage of data analysis is collection of data and in this analysis the dataset has been collected from

“<https://www.kaggle.com/datasets/prasad22/healthcare-dataset>”. The entire dataset consists of s 55,500 records for 15 variables and it may contain any type of data about the patients, their diseases, and organizing medical work.



**Figure 1: Data description**

(Source: Acquired from google colab)

Every record has essential details like Patient name, age and gender, patients’ blood group, medical history, medication history, test reports and other reports necessary for the treatment, doctor’s name, hospital name, Insurance company, type of admission, room number and amount to be billed. Admission and discharge dates also feature into the dataset and are important when evaluating the number of days any given patient had been in the hospital. From self-reported age the respondents can be aged between 13 and 89 years While from the self-reported billing amount, they ranged from negative values to $50,559. The range of the room numbers starts inclusive of 101 and ends at 500, an even distribution over various areas of the hospital.

## Proposed cloud based solution

In order to meet the requirements necessary for efficient management of healthcare data, and tools for analysis of this data, the cloud-based solution is suggested. However, with the benefits that come with cloud based infrastructure such as elasticity, backup and cost reduction, this solution is unique in that it is able to process data and deploy models with efficiency.

***Scalability***

The means by which the system can generate insightful results is also scalable; as the size of the dataset increases or the number of users actively seeking information concurrently rises, more processing and storage capacity can be provided through the cloud technology. That way the performance is maintained at optimum standards right from the onset as it does not require any form of intervention.

***Redundancy***

High availability is also possible by backing up a data center with a Redundancy plan through cloud storage solutions. Data is synchronized to physical machines, servers, or data centers, ensuring the solution for failing hardware components or data center outages. This ensures that important data which is often used in the healthcare process is always available and safe.

***Cost-Effectiveness***

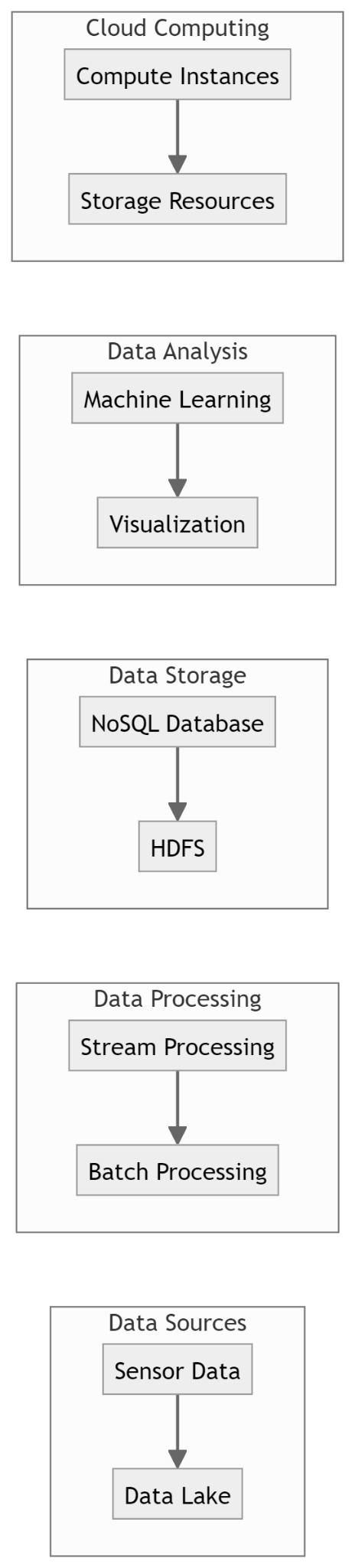
These multinational enterprise solutions have no upfront fixed costs as they use the utility computing billing structure where clients pay for the number of resources consumed. However, with cloud based solutions, the solution utilizes serverless computing for data preprocessing and machine learning activities in a manner that these are only activated when necessary. They do away with the need for having to constantly maintain expensive servers that are always online; this means cutting on costs.

## Data ingestion process

This simply means the ability to move healthcare data to cloud environments without much interference. However, by utilizing some services offered by cloud storage solution, data can be transferred from on-premises systems or from other sources to the cloud in a secure and efficient manner. This process provides multiple formats for data and offers auto, scheduled, and/or real-time transfer capabilities. These technologies help to maintain data accuracy, consistency, and privacy during the transfer phase, all of which are crucial prerequisites for further analysis in the cloud computing infrastructure.

# Scalable Processing Architecture

## Big data processing architecture



**Figure 2: Scalable Processing Architecture**

(Source: Self-created)

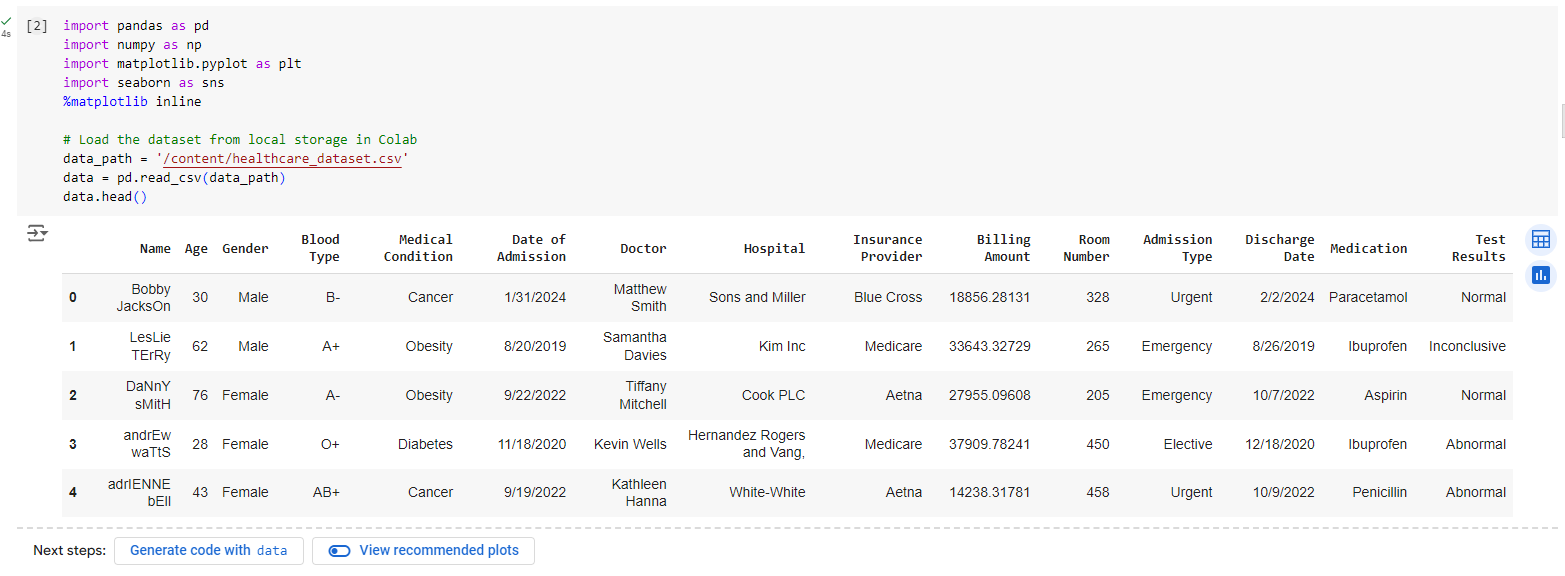
The above figure illustrates the entire architecture using the cloud computing resources.

## Choice of cloud services

In this study while considering distributed computing and parallel processing in the environments of big data, the proper choice of the cloud services is a critical and must-successful step in terms of scalability and cost. Various cloud platforms like AWS, GCP, Azure have their managed services in the form of EMR, Dataproc, HDInsight, etc, which provide Apache Spark, Hadoop, and other frameworks as pre-built clusters for ease of setup and management. These services are useful due to their ability to scale the clusters according to the complexity of workloads while at the same time, cutting infrastructure management costs. However, for the same reason, cloud storage providers also provide pay-as-you-go services, users need to pay for the amount of resource they use and thus cloud storage solutions are cost efficient. Therefore, adopting managed cloud services enhances distributed computing and parallel processing without inconveniences caused by scalabilities in distributed systems.

# Data Extraction and Pre-processing

## Data extraction process



**Figure 3: Data Extraction**

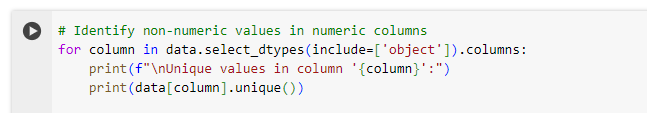
(Source: Acquired from Google colab)

The code illustrated represents actual usage of web scraping in Google Colab for extracting useful data. It starts by loading the required data manipulation tools and data visualization tools and libraries such as pandas, numpy, matplotlib and seaborn. The data set to be used in analysis is assumed to be stored in the current local Colab environment and is read using pandas. Sorting by the specifically selected columns is also allowed, so the set of basic information allows one to know the characteristic of the structure that includes the first few rows of the dataset and its specific summary statistics. Prediction for missing values- this is done by imputing them with the mean for continuous variables. Data analysis tools encompass conversion of data into histograms that give an overview of the data distribution and correlation heat maps that showcase the relationships between two variables (Petracchi *et al.* 2024). Extra treatments, including feature conversion for categorical inputs and data storage in raw and processed formats, are coded in a commented section for either inclusion or exclusion. This process helps in carrying out exhaustive evaluation and pre-processing of the data before engaging in further operations inside the modeling pipeline.

In the data extraction process the data or information is first collected from various datasets or databases. In this case, debitage data is extracted from a local file in Google Colab using the Pandas library. Ordering non-numeric values in numeric columns is detected and handled by converting the NaN values to numeric using pd. to\_numeric(). It also deals with cases when the value is missing and if the dataset includes categorical data, they are further encoded. These transformations help to maintain consistency as well as integrity of the data. The process makes the data easy to structure for future use to allow a proper level of analysis and decision making based on accurate insights.

**Data cleaning**

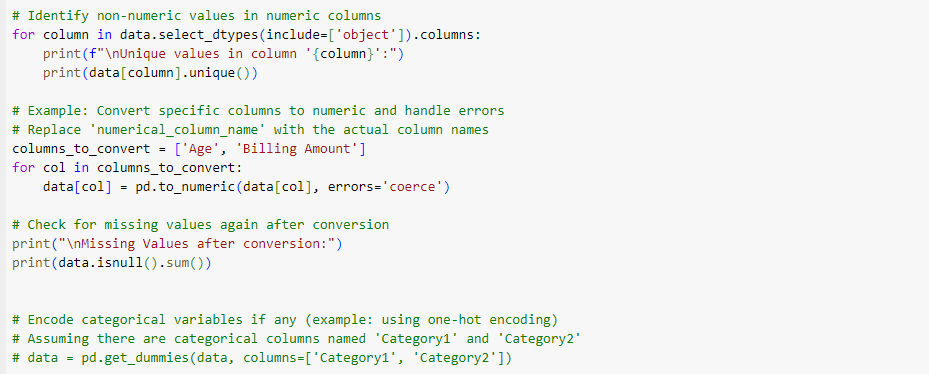
Data preprocessing has been known to be another key step to perform in the process of data analysis, as there has been a possibility of having an unstructured dataset.

****

**Figure 4: Identifying non-numeric values**

(Source: Acquired from Google colab)

The above figure describes the checking for non-numeric values that might have been entered and which should not be allowed in numerate columns. Therefore, for each numeric column, all of the unique values in that column are presented for the user’s review.



**Figure 5: Data cleaning**

(Source: Acquired from Google colab)

The above code snippet describes that, identifying and converting object values has been done in numeric fields into numeric types to make the data more consistent. It then looks for null values after converting and if so, it proceeds to codify categorical data using one hot encoding when necessary. This process ensures that the data types of the categorical variables being used are consistent and suited for the next step of data preparation.

# Data Analysis and Insights

## Cloud-based analytics solution for analyzing

Analytics solutions that leverage cloud infrastructures for analysis of healthcare data have been initiated by the sourcing of data from one or many datasets or databases. The loaded data is in the Cloud environment and this is done using tools such as Pandas. Notation of strings in a numeric format and outliers are detected and handled, as the complete dataset is constructed, and if the dataset has any categorical variables, these are also encoded (GIWA, 2024). These transformations help to ensure that the data is consistent and accurate across the data warehouse. However, it becomes possible to utilize advanced analytics tools and different techniques, including an intelligent agent’s machine learning model, to perform analyses and make decisions. Cloud computing is suitable for healthcare data processing as server management is not a burden for data processing when it is outsourced in the cloud, and the data grows exponentially with people’s population increases.

## Machine learning algorithms applied

The machine learning techniques or algorithms applied in this entire data analysis process has been discussed in this section. However, algorithms from machine learning used include RandomForestClassifier and is sourced from scikit-learn.



**Figure 6: Machine learning algorithms**

(Source: Acquired from Google colab)

In order to make the analysis, the dataset is divided into training and testing sets using Keras inbuilt function train\_test\_split. The learned model is reinitialized and optimized on the provided training data. It is chosen that predictions are made on the testing data and accuracy is elaborated by using accuracy\_score (Kunduru, 2023). This way, one can employ predictive modeling to categorize data by its features, having the accuracy rate as a reception parameter.

## Analysis contribution to informed decision-making

This is a very beneficial approach to identify specific insights from healthcare data so the machine learning algorithms can be used to make data decisions in the enterprise. Consequently, classification algorithms such as RandomForestClassifier in the context of a forecasting model show the probabilities of risk, diagnosis, and planning of further treatment or health resource management. This rational decision making, thus improves the delivery of healthcare, the hospital care people receive, and ultimately the organizational effectiveness within the enterprise.

# Cost Optimization Strategies

## Strategies for optimizing costs associated

In order to reduce the cost on the storage and computational resources required for the data on cloud, some measures include data segmentation for high/ low frequency access, use of serverless architectures, and auto-scaling of resources for on-demand processing. Moreover, employ the cost saving measures that are offered by cloud service providers to track down on cost wastage and readjust the ongoing use of resources.

Efficiency of resources can be a major step achieved through use of auto-scaling facilities to align provision of computational resources with demands for efficiency. Light utilities are another useful feature; a form of pre-built instances, reserved instances can provide up to 63% of the price discount based on the customer’s consolidated usage compared to on-demand instances. Amazon S3 has an example of its infrequent storage access storage class or Google Cloud Storage Nearline for cheaper storage costs of data that is less frequently accessed. However, through these strategies, organizations can thus be in a position to control cost while at the same time ensuring that they are able to meet the necessary levels of actual performance and bandwidth required for their data processing needs in the cloud.

# Security and Compliance

## Security measures and compliance considerations

In this entire analysis it is still pertinent for sensitive health care information that security measures and compliance issues are well observed when using the cloud. This involves accessing the database, files, and transmission of data with an aim of protecting the content as it is stored and when in the process of being transferred between endpoint devices. Development strategies of passing codes to control the incoming traffic Access control mechanisms and encryption practices has been taken.

# Performance Monitoring and Management

## Outline strategies for monitoring the performance of big data

Some strategies for big data performance measurement include monitoring of resource consumption, job performance metrics, throughput rates, data and process quality, and measuring scalability. These aspects assist monitor resources, facilitate identification of possible constraints in the system, and can guarantee information accuracy and adaptability to different workloads.

## Exploring tools

Loading and preprocessing of the data has been done using the pandas library and get visualizations and data analysis from the matplotlib and seaborn libraries and implement the machine learning algorithms with the help of scikit-learn library. For these reasons, it is possible and easy to use all of these tools on Google Colab for an efficient analysis of the data as well as incorporating them in a manner that allows for easy execution of data exploration and analysis tools.

# References

GIWA, Y.A., 2024. OVER THE COUNTER STOCKS DATA ACQUISITION & ANALYSIS WITH TIME SERIES PREDICTION. International Journal of Social Sciences and Scientific Studies, 4(1), pp.3643-3670.

Kunduru, A.R., 2023. Recommendations to advance the cloud data analytics and chatbots by using machine learning technology. International Journal of Engineering and Scientific Research, 11(3), pp.8-20.

Petracchi, E.J., Olivieri, S.E., Varela, J., Canullan, C.M., Zandalazini, H., Ocampo, C. and Quesada, B.M., 2024. Use of artificial intelligence in the detection of the critical view of safety during laparoscopic cholecystectomy. Journal of Gastrointestinal Surgery.