**Project Title: Big Data Analysis with IBM Cloud Databases**

**Phase 1: Problem Definition and Design Thinking**

In this part you will need to understand the problem statement and create a document on what have you understood and how will you proceed ahead with solving the problem. Please think on a design and present in form of a document.

**Problem Definition:**The project involves delving into big data analysis using IBM Cloud Databases. The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns. The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and visualizing the results for business intelligence.

**Design Thinking:**

**1. Data Selection**: Identify the datasets to be analyzed, such as climate data or social media trends.

**2. Database Setup**: Set up IBM Cloud Databases for storing and managing large datasets.

**3. Data Exploration**: Develop queries and scripts to explore the datasets, extract relevant information, and identify patterns.

**4. Analysis Techniques**: Apply appropriate analysis techniques, such as statistical analysis or machine learning, to uncover insights

**5. Visualization:** Design visualizations to present the analysis results in an understandable and impactful manner.

**6. Business Insights:** Interpret the analysis findings to derive valuable business intelligence and actionable recommendations.

Phase 1 link:

<https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/CAD_Phase1.docx>

**Phase 2:Innovation**

In this section we need to put our design into innovation to solve the problem. We have to consider incorporating advanced machine learning algorithms for predictive analysis or anomaly detection in the big data.

**Algorithm used in big data analysis:**

**Logistic Regression:**

Logistic regression is a [supervised machine learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) algorithm mainly used for [classification](https://www.geeksforgeeks.org/getting-started-with-classification/) tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it’s referred to as regression because it takes the output of the [linear regression](https://www.geeksforgeeks.org/ml-linear-regression/)function as input and uses a sigmoid function to estimate the probability for the given class. The [difference between linear regression and logistic regression](https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/) is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

**Type of Logistic Regression:**

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
3. **Ordinal:**In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

**Visualization methods:**

### **Scatter Plot:**

Scatter plots are used to observe relationships between variables and uses dots to represent the relationship between them. The [**scatter()**](https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/) method in the matplotlib library is used to draw a scatter plot.

### **Histogram:**

A [histogram](https://www.geeksforgeeks.org/plotting-histogram-in-python-using-matplotlib/) is basically used to represent data in the form of some groups. It is a type of bar plot where the X-axis represents the bin ranges while the Y-axis gives information about frequency. The **[hist()](https://www.geeksforgeeks.org/matplotlib-pyplot-hist-in-python/)** function is used to compute and create a histogram. In histogram, if we pass categorical data then it will automatically compute the frequency of that data i.e. how often each value occurred.

[**Line Chart**](https://www.geeksforgeeks.org/line-chart-in-matplotlib-python/) **:**

[Line Chart](https://www.geeksforgeeks.org/line-chart-in-matplotlib-python/) is used to represent a relationship between two data X and Y on a different axis. It is plotted using the **plot()** function.

**3D Scatter Plot** :

A **3D Scatter Plot** is a mathematical diagram, the most basic version of three-dimensional plotting used to display the properties of data as three variables of a dataset using the cartesiancoordinates.To create a 3D Scatter plot, Matplotlib’s **mplot3d toolkit** is used to enable three dimensional plotting.Generally 3D scatter plot is created by using ax.scatter3D() the function of the matplotlib library which accepts a data sets of X, Y and Z to create the plot while the rest of the attributes of the function are the same as that of two dimensional scatter plot.

**Pie chart:**

A pie chart is a pictorial representation of data in the form of a circular chart or pie where the slices of the pie show the size of the data. A list of numerical variables along with categorical variables is needed to represent data in the form of a pie chart. The arc length of each slice and consequently the area and central angle it forms in a pie chart is proportional to the quantity it represents.

### **Line Plot with seaborn:**

[Line Plot](https://www.geeksforgeeks.org/data-visualization-with-seaborn-line-plot/) in Seaborn plotted using the **[lineplot()](https://www.geeksforgeeks.org/seaborn-lineplot-method-in-python/)** method.  In this, we can pass only the data argument also.

**Contour plot:**

A contour plot is a graphical technique for representing a 3-dimensional surface by plotting constant z slices, called contours, on a 2-dimensional format. That is, given a value for z, lines are drawn for connecting the (x,y) coordinates where that z value occurs.

**Steps to be followed in the cloud database :**

**Step 1: Data Collection and Preparation**

**Data Sources Identification**: Identify the sources of the data you plan to analyze. This could include internal databases, external APIs, or third-party data providers.

**Data Extraction and Integration:** Extract data from various sources and integrate it into a unified format suitable for analysis.

**Data Cleaning:** Cleanse the data to handle missing values, outliers, and inconsistencies.

Data Transformation: Prepare the data for analysis by performing tasks like normalization, aggregation, or feature engineering.

**Step 2: IBM Cloud Setup**

**IBM Cloud Account Creation:** If not already done, create an IBM Cloud account.

**Select Database Service:** Choose the appropriate IBM Cloud Database service based on your data requirements (e.g., IBM Db2, IBM Cloudant, or IBM Db2 on Cloud).

**Provision Database:** Set up the database instance on IBM Cloud, configure security settings, and ensure proper access control.

**Step 3: Data Loading and Storage**

**Data Ingestion:**Load the prepared data into the IBM Cloud Database. Depending on the volume, you may need to implement data ingestion pipelines.

**Data Security:**Implement encryption and access controls to protect sensitive data.

**Data Backup and Recovery:** Establish backup and recovery mechanisms to safeguard data.

**Dataset selection:**

I am select the database consisting of Customer id,Total\_bill,tip,sex,smoker,day,time,size and phone number and further details about the customers of the hotel.

**Dataset Link:**

<https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/Customers-200.csv>

**Libraries used in ibm cloud database(python):**

importnumpy as np

import pandas as pd

frommatplotlib import pyplot as plt

importseaborn as sns

importplotly.express as px

fromsklearn.model\_selection import train\_test\_split

fromsklearn.preprocessing import StandardScaler

fromsklearn.linear\_model import LogisticRegression

fromsklearn.metrics import confusion\_matrix

fromsklearn.metrics import accuracy\_score

# Steps in cloud database to work with datasets:

1. Download the CSV files from any websites.
2. Start a New Database **Database -> New Database**
3. Start the import **Database -> Import**
4. Select the file to import (start with articles.csv)
5. Give the table a name that matches the file name (articles, journals, licences, languages publishers), or use the default
6. Since the first row has column headings, check the “First row contains column names”- box
7. Under “Fields separated by”, check “Comma”. Ensure ‘Ignore trailing Separator/Delimiter’ is left *unchecked*.
8. Also, under “Fields enclosed by”, ensure that “Double quotes if necessary” is left *checked*.
9. Press **OK**
10. When asked if you want to modify the table, click **OK**
11. Set the data types for each field and INTEGER for fields with numbers:
12. Click **OK**

Phase 2 link:

<https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/CAD_Phase2.docx>

**Phase 3: Development Part 1**

* In this part you will begin building your project.
* Start building the big data analysis solution using IBM Cloud Databases.
* Create an IBM Cloud account, choose the appropriate database service (e.g., Db2, MongoDB), and set up a database instance.
* Develop queries or scripts to explore and analyze the selected dataset. Perform basic data cleaning and transformation as needed.

**Phase3 link:**

[**https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/CAD\_Phase3.docx**](https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/CAD_Phase3.docx)

**The coding is given in the below link:**

<https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/Big_Data_Analysis_Part1.ipynb>

**Big Data Analysis WithIbm Cloud Databases**

**Phase 3: Development Part 2**

In this part you will continue building your project.

* Continue building the big data analysis solution by applying advanced analysis techniques and visualizing the results.
* Apply more complex analysis techniques, such as machine learning algorithms, time series analysis, or sentiment analysis, depending on the dataset and objectives.
* Create visualizations to showcase the analysis results. Use tools like Matplotlib, Plotly, or IBM Watson Studio for creating graphs and charts.

**Phase 4 link:**

<https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/CAD_Phase4.docx>

**The coding is given in the below link:**

<https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/Big_Data_Analysis_Part2.ipynb>

**The readme file gives a detailed information about the Big Data Analysis in a dataset by logistic regression and usingsome visualization methods,**

[**https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/README.md**](https://github.com/Subanandhini-15/Naan-Mudhalvan/blob/main/README.md)