# Big Data Analysis with IBM Cloud Databases for Iris Dataset

## Problem Definition

## A miniature but ubiquitous dataset for testing machine learning techniques and data visualizations is called the iris dataset. It consists of 150 observations on the sepal length, sepal breadth, petal length, and petal width of three different kinds of iris flowers (setosa, versicolor, and virginica). The goal of this project is to manage and maintain the iris dataset using IBM Cloud Databases while also doing various data analysis jobs to gain useful insights from the data. Using the necessary visualizations and business intelligence tools, the project will also involve developing and presenting the analytical results.

## Design Thinking

The following steps will be followed to complete the project:

### Data Selection

### For this research, the iris dataset will serve as the primary data source. The five columns in the dataset are species, petal length, petal width, and sepal width. The variable with the goal value that identifies the species of iris flower is the species column.

### Database Setup

IBM Cloud Databases will be used to store and manage the iris dataset in a scalable and secure way. IBM Cloud Databases provides various types of databases, such as relational, document, graph, and time series databases. For this project, a relational database will be used, as it is suitable for structured data with predefined schema. IBM Cloud Databases also provides tools for data integration, backup, monitoring, and performance optimization.

To set up a relational database using IBM Cloud Databases, the following steps will be performed:

* Create an IBM Cloud account and log in to the IBM Cloud console.
* Navigate to the Catalog and search for “Databases”.
* Select “Databases for PostgreSQL” as the database service.
* Configure the service name, region, resource group, plan, and other settings as per the requirements.
* Click on “Create” to provision the database service.
* Once the service is created, click on “Service Credentials” and generate a new credential.
* Copy the connection string and other details from the credential.
* Use a database client tool, such as pgAdmin or DBeaver, to connect to the database using the connection string and credential details.
* Create a new database schema and a new table for the iris dataset.
* Import the iris dataset from the CSV file into the table using the database client tool or a command line tool.

### Data Exploration

After setting up the database and importing the data, the next step is to explore the data and perform some preliminary analysis. This step will involve developing queries and scripts to access and manipulate the data using SQL or other languages. Some of the tasks that will be performed in this step are:

* Check the basic information about the data, such as number of rows, columns, data types, missing values, duplicates, etc.
* Calculate some descriptive statistics for each feature and target variable, such as mean, median, mode, standard deviation, range, etc.
* Perform some data cleaning and preprocessing tasks, such as handling missing values, outliers, errors, etc.
* Perform some data transformation tasks, such as scaling, normalization, encoding, etc.
* Perform some feature engineering tasks, such as creating new features from existing ones, selecting relevant features, etc.

### Analysis Techniques

After exploring and preparing the data, the next step is to apply appropriate analysis techniques to uncover insights from the data. This step will involve using various statistical methods or machine learning algorithms to perform different types of analysis tasks. Some of the tasks that will be performed in this step are:

* Perform some exploratory data analysis (EDA) tasks to understand the distribution and relationship of the data using visual techniques such as histograms, box plots, scatter plots, etc.
* Perform some inferential statistics tasks to test some hypotheses or assumptions about the data using techniques such as t-test, ANOVA, chi-square test, correlation, regression, etc.
* Perform some predictive analytics tasks to build models that can predict the target variable (species) from the features using techniques such as classification, clustering, decision trees, random forests, k-nearest neighbors, support vector machines, neural networks, etc.
* Perform some prescriptive analytics tasks to provide recommendations or suggestions based on the analysis results using techniques such as optimization, simulation, scenario analysis, etc.

### Visualization

After performing various analysis tasks on the data, the next step is to design visualizations that can present the analysis results in an understandable and impactful manner. This step will involve using various tools and libraries to create charts, graphs, maps, dashboards, reports, etc. that can communicate the insights effectively. Some of the tools and libraries that will be used in this step are:

* Matplotlib: A Python library for creating static 2D plots.
* Seaborn: A Python library for creating statistical and aesthetic plots.
* Plotly: A Python library for creating interactive and web-based plots.
* Dash: A Python framework for building web applications and dashboards.
* IBM Cognos Analytics: A business intelligence tool for creating reports and dashboards.

### Business Insights

The final step of the project is to interpret the analysis findings and derive valuable business intelligence and actionable recommendations. This step will involve using the visualizations and reports to answer some business questions or solve some business problems related to the iris dataset. Some of the questions or problems that will be addressed in this step are:

* What are the main characteristics and differences of each iris species?
* How can the iris species be classified or clustered based on their features?
* How accurate are the predictive models for identifying the iris species?
* What are the best features or parameters for predicting the iris species?
* How can the analysis results be used to improve the quality or efficiency of iris cultivation or classification?