Lab 5 – Data Storage in the Cloud

Student: Ruth Ihunanya Chimezuru Obere
Course: Big Data Analytics_Lab 5
Dataset: Iris Dataset (`iris.csv`)

Part A: Creating a Cloud Storage Bucket

For this lab, I first created a Google Cloud Storage bucket to store my dataset.

Steps I followed:

1. I opened the **Google Cloud Console** and navigated to **Storage** → **Browser**.
2. I clicked **+ CREATE BUCKET** to create a new bucket.
3. I named my bucket **`ruth-lab5-data`**, making sure it was unique.
4. I selected the **Region** location type and chose a region in Europe.
5. I set the **default storage class** to **Standard**, which is suitable for frequently accessed data.
6. I chose **Fine-grained access control** to allow permissions for individual objects.
7. I clicked **CREATE**.

After the bucket was created, I uploaded my dataset (`iris.csv`) by clicking **UPLOAD FILES** and selecting the file from my system.

Part B: Accessing and Exploring Data from Jupyter Notebook

I used **Method 1: gsutil to Copy Files** to get the dataset into my Jupyter Notebook.

1. Copied the file from the bucket to my VM:

```
!gsutil cp gs://ruth-lab5-data/iris.csv .
```

2. Loaded the dataset into a Pandas DataFrame:

```
import pandas as pd

df = pd.read_csv("iris.csv")
```

3. Explored the dataset:

- **Check the first few rows:**

```
df.head()
```

- **Check number of rows and columns:**

```
df.shape
```

- **Display column names:**

```
df.columns
```

- **Understand data types and missing values:**

```
df.info()
```

- **View basic statistics:**

```
df.describe()
```

- **Check for missing values:**

```
df.isnull().sum()
```

- **Count samples per species:**

```
df['species'].value_counts()
```

The dataset contains 150 rows and 5 columns.
There are no missing values.
Each species (`setosa`, `versicolor`, `virginica`) has 50 samples.

Summary

- Successfully created a **cloud storage bucket** and uploaded the dataset.
- Accessed the dataset from **Jupyter Notebook** running on a cloud VM.
- Explored the dataset, checked data types, missing values, basic statistics, and distribution of species.
- Saved the notebook as Lab5_PartB_iris.ipynb