**PHASE - 3**

**BIG DATA ANALYSIS WITH IBM CLOUD DATABASES**

**INTRODUCTION:**

This phase is based on loading and preprocessing the dataset.In the realm of big data analysis, the foundation of success lies in the effective loading and preprocessing of datasets. When harnessed with IBM Cloud Databases, this process becomes a cornerstone for powerful insights. This introduction delves into the critical initial steps of handling vast datasets, ensuring they are optimized for analysis. Whether it's structured or unstructured data, real-time streams, or historical records, these procedures are essential for unlocking the full potential of IBM's robust database solutions. Prepare to embark on a data-driven journey that leads to profound insights with IBM Cloud Databases.

**LOADING AND PREPROCESSING THE DATASET:**

**STEP 1 :** The dataset is load into the model in the name of ‘bist100’ as a csv file.

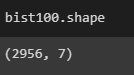


**STEP 2 :** Different analysis using the dataset.

**STEP 2.1 :** Casual analysis.

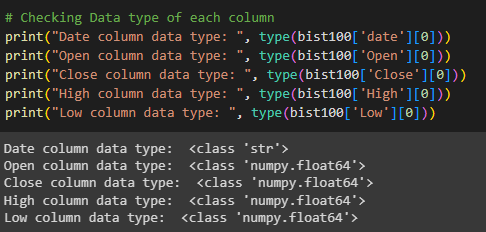
**STEP 2.1.1 :** Dimension of the dataset.

**Dimension:** The number of attributes/features that exist in a dataset.



**STEP 2.1.2 :** Determine the datatype of each column .

**Datatypes:** Data types are the classification or categorization of data items.For example string ,integer etc.

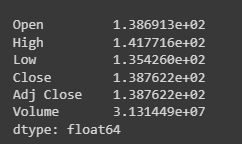


**STEP 2.2 :** Descriptive analysis.

**STEP 2.2.1 :** Calculating the mean for the all columns of the dataset.

**Mean:** The average of a data set, found by adding all numbers together and then dividing the sum of the numbers by the number of numbers.

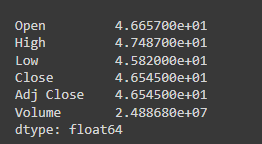




**STEP 2.2.2 :** Calculating the median for the all columns of the dataset.

**Median:** The median is the middle value of a set of numbers.

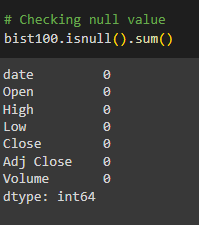




**STEP 2.3 :** Data cleansing

**STEP 2.3.1 :** Checking for the null values in all columns of the dataset.

**Null values:** A null value in a relational database is used when the value in a column is unknown or missing.



**STEP 2.3.2 :** Boolean method for calculating the null values for all columns of the dataset.

**Null values in Boolean :** The variable has no reference assigned, so it is neither true nor false, it is “nothing”

