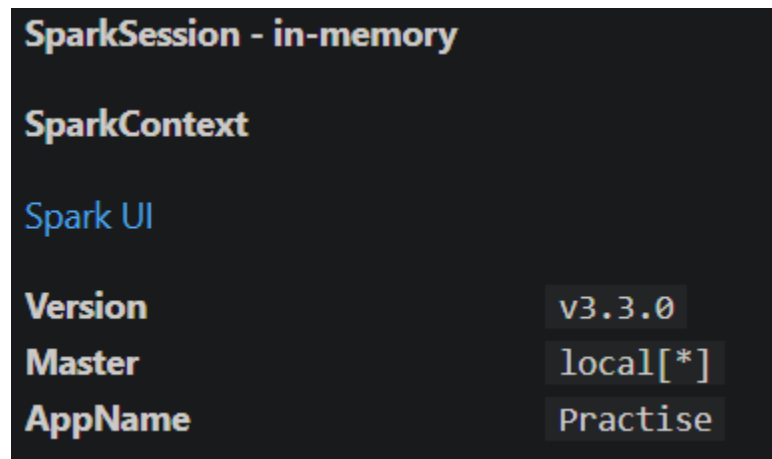


- Used for process huge amount of data
- **PySpark**
  - First of all we need to start a Spark session

```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('Practise').getOrCreate()
```

- When we are working in a cloud we can create multiple clusters
- When working with multiple instances we will see masters and cluster one, two , three etc



## - Part 1

- PySpark Dataframe
- Reading The Dataset
- Checking the Datatypes of the Column(Schema)
- Selecting Columns And Indexing
- Check Describe option similar to Pandas
- Adding Columns
- Dropping columns
- Renaming Columns

**InferSchema = True** ( By default pyspark read values as strings to avoid that we have to use **InferSchema = True**)

**What is a data frame**

- It is a kind of data structure

- **Part 2**

## Pyspark Handling Missing Values

- Dropping Columns
- Dropping Rows
- Various Parameter In Dropping functionalities
- Handling Missing values by Mean, MEdian And Mode

- **Part 3**

## Pyspark Dataframes

- Filter Operation
- $\&, |, =$
- $\sim$

- Applying multiple operations on a dataframe

- **Part 4**

## Pyspark GroupBy And Aggregate Functions

- First we have to use the GroupBy function and after that we have to apply Aggregate functions.

- **Part 5**

There are 2 different technologies in Spark ML

1. RDD Techniques
2. Data frame APIs ( Recent one and most used one)