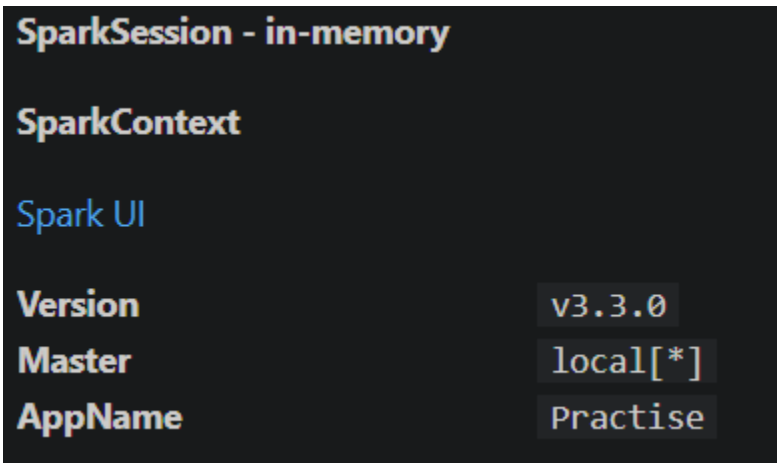
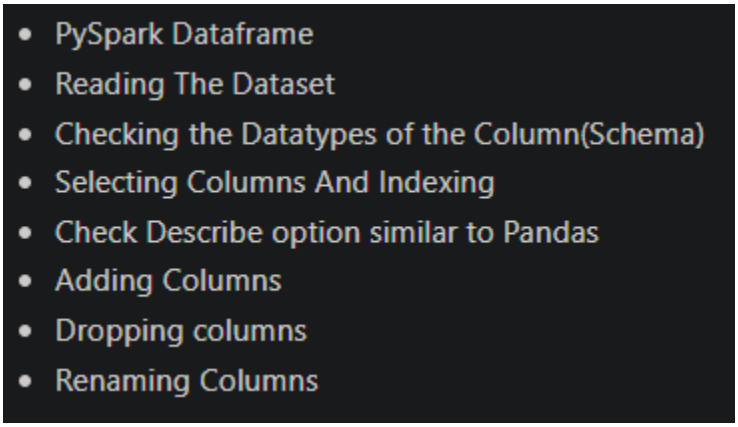
- Used for process huge amount of data
- **PySpark**
    - First of all we need to start a Spark session

      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**

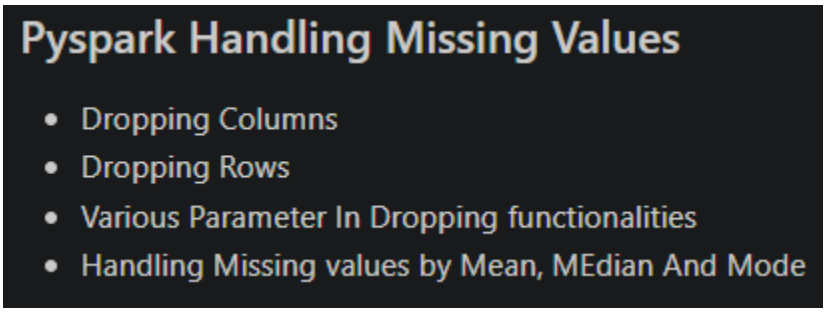


InferSchema = True (By default pyspark read values as strings to avoid that we have to use InferSchema = True)

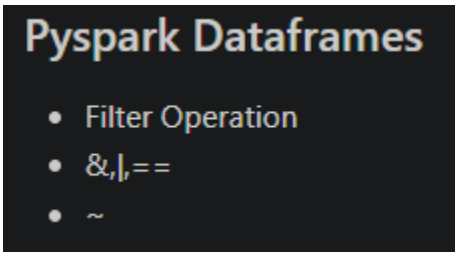
What is a data frame

- **It is a kind of data structure**

- **Part 2**

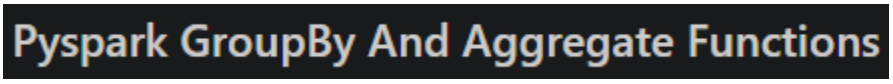


- **Part 3**



  - Applying multiple operations on a dataframe

- **Part 4**



  - **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)