In [5]:

```
!pip install pyspark

Apache Spark is unified analytics engine for large scale data processing
```

```
Requirement already satisfied: pyspark in c:\users\aviral\appdata\local\programs\pyth
on\python311\lib\site-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in c:\users\aviral\appdata\local\progra
ms\python\python311\lib\site-packages (from pyspark) (0.10.9.7)
```

In [6]:

```
import pyspark
```

In [25]:

```
import pandas as pd
pd.read_csv('D:\Data Analysis\Python\python projects\Test.csv')
##type(pd.read_csv('D:\Data Analysis\Python\python projects\Test.csv'))
```

Out[25]:

	Name	Age
0	Nishant	33
1	Neha	30
2	Aviral	24
3	Kanti	55

In [12]:

```
from pyspark.sql import SparkSession


#SparkSession.builder: This initializes the builder for creating a SparkSession.
#appName('DataAnalysisTesting'): This sets the name of your Spark application to "Data
#getOrCreate(): This method either returns an existing SparkSession or creates a new c
```

In [17]:

```
spark=SparkSession.builder.appName('D:\Data Analysis\Python\python projects\Testing').
```

In [18]:

```
spark
```

Out[18]:

SparkSession - in-memory

SparkContext

Spark UI

Version	`v3.5.1`
Master	`local[*]`
AppName	`D:\Data Analysis\Python\python projects\Testing`

In [19]:

```
df_pyspark=spark.read.csv('D:\Data Analysis\Python\python projects\Test.csv')
```

In [20]:

```
df_pyspark.show()
```

```
+-------+---+
|    _c0|_c1|
+-------+---+
|   Name|Age|
|Nishant| 33|
|   Neha| 30|
| Aviral| 24|
|  Kanti| 55|
+-------+---+
```

In [ ]:

```
spark.read.option('header','true').csv('D:\Data Analysis\Python\python projects\Test.c

#spark.read.csv: This reads a CSV file and returns a DataFrame.
#header=True: This parameter indicates that the first row of the CSV file contains col
#inferSchema=True: This tells Spark to automatically infer the data types of the colum
```

In [27]:

```
type(df_pyspark)
```

Out[27]:

```
pyspark.sql.dataframe.DataFrame
```

In [29]:

```
df_pyspark.head(4)
```

Out[29]:

```
[Row(_c0='Name', _c1='Age'),
 Row(_c0='Nishant', _c1='33'),
 Row(_c0='Neha', _c1='30'),
 Row(_c0='Aviral', _c1='24')]
```

In [ ]: