Pyspark -

-Introduction -Spark session installation -Reading file and creating Pyspark dataframes -Reading the Dataset -Checking the datatypes of the columns(schema) -Selecting columns and indexing -Check describe option similar to pandas -Adding Columns -Dropping columns -Renaming columns

In [2]:

```
pip --proxy http://[username]:[password]@noidaproxy.corp.exlservice.com:8000 install pyspark
```

```
Collecting pyspark
  Downloading https://files.pythonhosted.org/packages/b8/01/b2393cee7f6180d91
50274e92c8bdc1c81220e2ad7554ee5febca1866899/pyspark-3.3.0.tar.gz (281.3MB)
Collecting py4j==0.10.9.5 (from pyspark)
  Using cached https://files.pythonhosted.org/packages/86/ec/60880978512d5569
ca4bf32b3b4d7776a528ecf4bca4523936c98c92a3c8/py4j-0.10.9.5-py2.py3-none-any.w
hl
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py): started
  Building wheel for pyspark (setup.py): still running...
  Building wheel for pyspark (setup.py): finished with status 'done'
  Stored in directory: C:\Users\shrinath195156\AppData\Local\pip\Cache\wheels
\9e\c1\93\d40ec851fc2b278e1056c1353ff95a7a4ef1b219f74ca9c11f
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.3.0
Note: you may need to restart the kernel to use updated packages.
```

In [1]:

```
import pyspark
```

In [2]:

```
#reading dataset using pandas
import pandas as pd
pd.read_csv("sparktest.csv")
```

Out[2]:

	name	age
0	Krish	31
1	Sudhanshu	30
2	Sunny	29

In [3]:

```
from pyspark.sql import SparkSession
```

In [4]:

```
#creating variable and session name
spk = SparkSession.builder.appName("Practice").getOrCreate()
```

In [5]:
```
spk
```

Out[5]:

SparkSession - in-memory

SparkContext

[Spark UI (http://EXLAPLPNyCdxfzp.corp.exlservice.com:4040)](http://EXLAPLPNyCdxfzp.corp.exlservice.com:4040)

Version

`v3.3.0`

Master

`local[*]`

AppName

`Practice`

In [7]:
```
#reading dataset using spark
df_pyspark=spk.read.csv("sparktest.csv")
df_pyspark
```

Out[7]:
```
DataFrame[_c0: string, _c1: string]
```

In [9]:
```
#view entire dataset with default headers values
df_pyspark.show()
```

```
+---------+---+
|      _c0|_c1|
+---------+---+
|     name|age|
|    Krish| 31|
|Sudhanshu| 30|
|    Sunny| 29|
+---------+---+
```

In [13]:
```
#calling in actual data headers
spk.read.option("header","true").csv("sparktest.csv")
```

Out[13]:
```
DataFrame[name: string, age: string]
```

In [14]:
```
#displaying entire data with headers
spk.read.option("header","true").csv("sparktest.csv").show()
```

```
+---------+---+
|     name|age|
+---------+---+
|    Krish| 31|
|Sudhanshu| 30|
|    Sunny| 29|
+---------+---+
```

In [22]:
```
df_pyspark = spk.read.option("header","true").csv("sparktest.csv")
```

In [23]:
```
#type of dataframe
type(df_pyspark)
```

Out[23]:
```
pyspark.sql.dataframe.DataFrame
```

In [24]:
```
#head view of dataframe
df_pyspark.head(3)
```

Out[24]:
```
[Row(name='Krish', age='31'),
 Row(name='Sudhanshu', age='30'),
 Row(name='Sunny', age='29')]
```

In [26]:
```
#print Schema works as df.info from pandas
df_pyspark.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- age: string (nullable = true)
```

In [2]:
```
from pyspark.sql import SparkSession
```

In [3]:
```
spark=SparkSession.builder.appName("Dataframe_Processing").getOrCreate()
```

In [4]:
```
spark
```

Out[4]:

SparkSession - in-memory

SparkContext

[Spark UI (http://EXLAPLPNyCdxfzp.corp.exlservice.com:4040)](http://EXLAPLPNyCdxfzp.corp.exlservice.com:4040)

Version

`v3.3.0`

Master

`local[*]`

AppName

`Dataframe_Processing`

In [5]:
```
#Reading dataset using option
spark.read.option("header","true").csv("sparktest.csv").show()
```

```
+---------+---+----------+
|     name|age|experience|
+---------+---+----------+
|    Krish| 31|         5|
|Sudhanshu| 30|         8|
|    Sunny| 29|         6|
+---------+---+----------+
```

In [7]:
```
#Checking the schema
df_pyspark = spark.read.option("header","true").csv("sparktest.csv")
df_pyspark
```

Out[7]:
```
DataFrame[name: string, age: string, experience: string]
```

In [8]:
```
#Checking the schema using inferschema for the non string values e.g. age in above result
df_pyspark = spark.read.option("header","true").csv("sparktest.csv",inferSchema=True)
df_pyspark
```

Out[8]:
```
DataFrame[name: string, age: int, experience: int]
```

In [9]:
```
df_pyspark.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- experience: integer (nullable = true)
```

In [63]:
```
#Reading dataset using read
df_pyspark = spark.read.csv("sparktest.csv", header=True, inferSchema=True)
df_pyspark.show()
```

```
+---------+---+----------+
|     name|age|experience|
+---------+---+----------+
|    Krish| 31|         5|
|Sudhanshu| 30|         8|
|    Sunny| 29|         6|
+---------+---+----------+
```

In [64]:
```
df_pyspark.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- experience: integer (nullable = true)
```

In [14]:
```
type(df_pyspark)
```

Out[14]:
```
pyspark.sql.dataframe.DataFrame
```

In [15]:
```
#getting columns
df_pyspark.columns
```

Out[15]:
```
['name', 'age', 'experience']
```

In [16]:
```
df_pyspark.head(3)
```

Out[16]:
```
[Row(name='Krish', age=31, experience=5),
 Row(name='Sudhanshu', age=30, experience=8),
 Row(name='Sunny', age=29, experience=6)]
```

In [17]:
```
#display dataframe
df_pyspark.show()
```

```
+---------+---+----------+
|     name|age|experience|
+---------+---+----------+
|    Krish| 31|         5|
|Sudhanshu| 30|         8|
|    Sunny| 29|         6|
+---------+---+----------+
```

In [20]:
```
#selecting data from only one column
df_pyspark.select("name").show()
```

```
+---------+
|     name|
+---------+
|    Krish|
|Sudhanshu|
|    Sunny|
+---------+
```

In [21]:
```
type(df_pyspark.select("name"))
```

Out[21]:
```
pyspark.sql.dataframe.DataFrame
```

In [22]:
```
#selecting data from multiple columns
df_pyspark.select(["name","experience"]).show()
```

```
+---------+----------+
|     name|experience|
+---------+----------+
|    Krish|         5|
|Sudhanshu|         8|
|    Sunny|         6|
+---------+----------+
```

In [24]:
```
#another way to select a column name
df_pyspark["name"]
```

Out[24]:
```
Column<'name'>
```

In [27]:
```
df_pyspark.dtypes
```

Out[27]:
```
[('name', 'string'), ('age', 'int'), ('experience', 'int')]
```

In [29]:

```
#Getting dataframe statistics using describe
df_pyspark.describe().show()
```

```
+-------+-----+----+------------------+
|summary| name| age|        experience|
+-------+-----+----+------------------+
|  count|    3|   3|                 3|
|   mean| null|30.0| 6.333333333333333|
| stddev| null| 1.0|1.5275252316519468|
|    min|Krish|  29|                 5|
|    max|Sunny|  31|                 8|
+-------+-----+----+------------------+
```

In [69]:

```
#addition of columns in dataframe using calculated column here
df_pyspark=df_pyspark.withColumn("experience after 2yrs",df_pyspark["experience"]+2)
```

In [73]:

```
df_pyspark.show()
```

```
+---------+---+----------+---------------------+
|     name|age|experience|experience after 2yrs|
+---------+---+----------+---------------------+
|    Krish| 31|         5|                    7|
|Sudhanshu| 30|         8|                   10|
|    Sunny| 29|         6|                    8|
+---------+---+----------+---------------------+
```

In [75]:

```
#dropping column from the dataframe have to use variable to pass the change
df_pyspark=df_pyspark.drop("experience after 2yrs")
```

In [76]:

```
df_pyspark.show()
```

```
+---------+---+----------+
|     name|age|experience|
+---------+---+----------+
|    Krish| 31|         5|
|Sudhanshu| 30|         8|
|    Sunny| 29|         6|
+---------+---+----------+
```

In [77]:

```
#renaming one column using with function
df_pyspark.withColumnRenamed("name", "First_Name").show()
```

```
+----------+---+----------+
|First_Name|age|experience|
+----------+---+----------+
|     Krish| 31|         5|
| Sudhanshu| 30|         8|
|     Sunny| 29|         6|
+----------+---+----------+
```

In [97]:

```
#renaming multiple required columns using with function
(df_pyspark.withColumnRenamed("name", "first_name")
 .withColumnRenamed("experience","total_exp")).show()
```

```
+----------+---+---------+
|first_name|age|total_exp|
+----------+---+---------+
|     Krish| 31|        5|
| Sudhanshu| 30|        8|
|     Sunny| 29|        6|
+----------+---+---------+
```

In [96]:

```
#another way of renaming all column names
refined_column_name_list = ["first_name","age","total_exp"]
df_pyspark1=df_pyspark.toDF(*refined_column_name_list).show()
```

```
+----------+---+---------+
|first_name|age|total_exp|
+----------+---+---------+
|     Krish| 31|        5|
| Sudhanshu| 30|        8|
|     Sunny| 29|        6|
+----------+---+---------+
```