Mod 1 - Overview - DataFrames

March 4, 2025

We will Cover

- 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

[1]:
```
from pyspark.sql import SparkSession
```

[2]:
```
spark=SparkSession.builder.appName('Dataframe').getOrCreate()
```

[5]:
```
spark
```

[5]:
```
<pyspark.sql.session.SparkSession at 0x7f0b0c363910>
```

[6]:
```
## read the dataset
df_pyspark=spark.read.option('header','true').csv('test1.csv',inferSchema=True)
```

[7]:
```
### Check the schema
df_pyspark.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

[8]:
```
df_pyspark=spark.read.csv('test1.csv',header=True,inferSchema=True)
df_pyspark.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Tiago| 31|        10| 30000|
|    Diogo| 30|         8| 25000|
|    Lucas| 29|         4| 20000|
|      Bia| 24|         3| 20000|
|Francisco| 21|         1| 15000|
|  Rodrigo| 23|         2| 18000|
+---------+---+----------+------+
```

```
[9]: ### Check the schema
     df_pyspark.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

```
[10]: type(df_pyspark)
```

```
[10]: pyspark.sql.dataframe.DataFrame
```

```
[11]: df_pyspark.head(3)
```

```
[11]: [Row(Name='Tiago', age=31, Experience=10, Salary=30000),
       Row(Name='Diogo', age=30, Experience=8, Salary=25000),
       Row(Name='Lucas', age=29, Experience=4, Salary=20000)]
```

```
[12]: df_pyspark.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Tiago| 31|        10| 30000|
|    Diogo| 30|         8| 25000|
|    Lucas| 29|         4| 20000|
|      Bia| 24|         3| 20000|
|Francisco| 21|         1| 15000|
|  Rodrigo| 23|         2| 18000|
+---------+---+----------+------+
```

```
[13]: df_pyspark.select(['Name','Experience']).show()
```

```
+---------+----------+
|     Name|Experience|
+---------+----------+
|    Tiago|        10|
|    Diogo|         8|
|    Lucas|         4|
|      Bia|         3|
|Francisco|         1|
|  Rodrigo|         2|
+---------+----------+
```

[14]:

```
df_pyspark['Name']
```

[14]:

```
Column<'Name'>
```

[15]:

```
df_pyspark.dtypes
```

[15]:

```
[('Name', 'string'), ('age', 'int'), ('Experience', 'int'), ('Salary', 'int')]
```

[16]:

```
df_pyspark.describe().show()
```

```
+-------+-----+------------------+-----------------+------------------+
|summary| Name|               age|       Experience|            Salary|
+-------+-----+------------------+-----------------+------------------+
|  count|    6|                 6|                6|                 6|
|   mean| NULL|26.333333333333332|4.666666666666667|21333.333333333332|
| stddev| NULL| 4.179314138308661|3.559026084010437| 5354.126134736337|
|    min|  Bia|                21|                1|             15000|
|    max|Tiago|                31|               10|             30000|
+-------+-----+------------------+-----------------+------------------+
```

[17]:

```
### Adding Columns in data frame
df_pyspark=df_pyspark.withColumn('Experience After 2␣
 ↪year',df_pyspark['Experience']+2)
```

[18]:

```
df_pyspark.show()
```

```
+---------+---+----------+------+-----------------------+
|     Name|age|Experience|Salary|Experience After 2 year|
+---------+---+----------+------+-----------------------+
|    Tiago| 31|        10| 30000|                     12|
|    Diogo| 30|         8| 25000|                     10|
|    Lucas| 29|         4| 20000|                      6|
|      Bia| 24|         3| 20000|                      5|
|Francisco| 21|         1| 15000|                      3|
|  Rodrigo| 23|         2| 18000|                      4|
+---------+---+----------+------+-----------------------+
```

[19]:

```
### Drop the columns
df_pyspark=df_pyspark.drop('Experience After 2 year')
```

[20]:

```
df_pyspark.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Tiago| 31|        10| 30000|
|    Diogo| 30|         8| 25000|
|    Lucas| 29|         4| 20000|
|      Bia| 24|         3| 20000|
|Francisco| 21|         1| 15000|
|  Rodrigo| 23|         2| 18000|
+---------+---+----------+------+
```

[21]:

```
### Rename the columns
df_pyspark.withColumnRenamed('Name','New Name').show()
```

```
+---------+---+----------+------+
| New Name|age|Experience|Salary|
+---------+---+----------+------+
|    Tiago| 31|        10| 30000|
|    Diogo| 30|         8| 25000|
|    Lucas| 29|         4| 20000|
|      Bia| 24|         3| 20000|
|Francisco| 21|         1| 15000|
|  Rodrigo| 23|         2| 18000|
+---------+---+----------+------+
```