

Name: Rohan Vinayak Chaudhari

Batch: Data Engineering

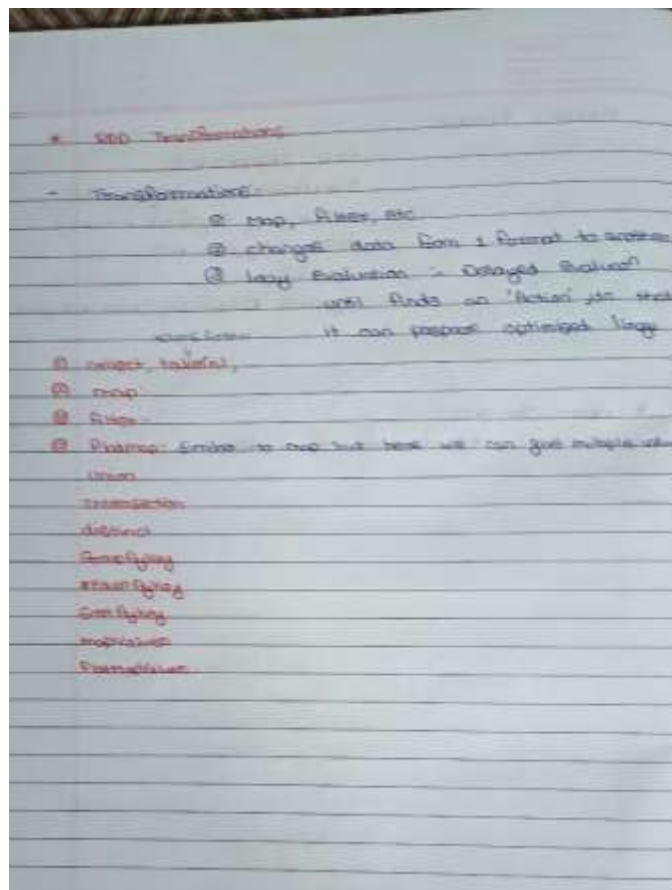
Date: 06/02/2024

Topic: Pyspark

Solution:

1. Pyspark:

Theory:



* DataFrame in PySpark

It is Data Structure

1) Read CSV

① spark.read.csv('filePath', -1 -)

② spark.read.option('header', 'true').csv('Path',
Here if we don't write this we consider
 the data as string
 inferSchema='true')

2) View head()

① Particular Column: <var>.select('col').show()

② <var>.select('col1', 'col2').show()

③ Datatype: <var>.dtype

④ Describe: <var>.describe().show()

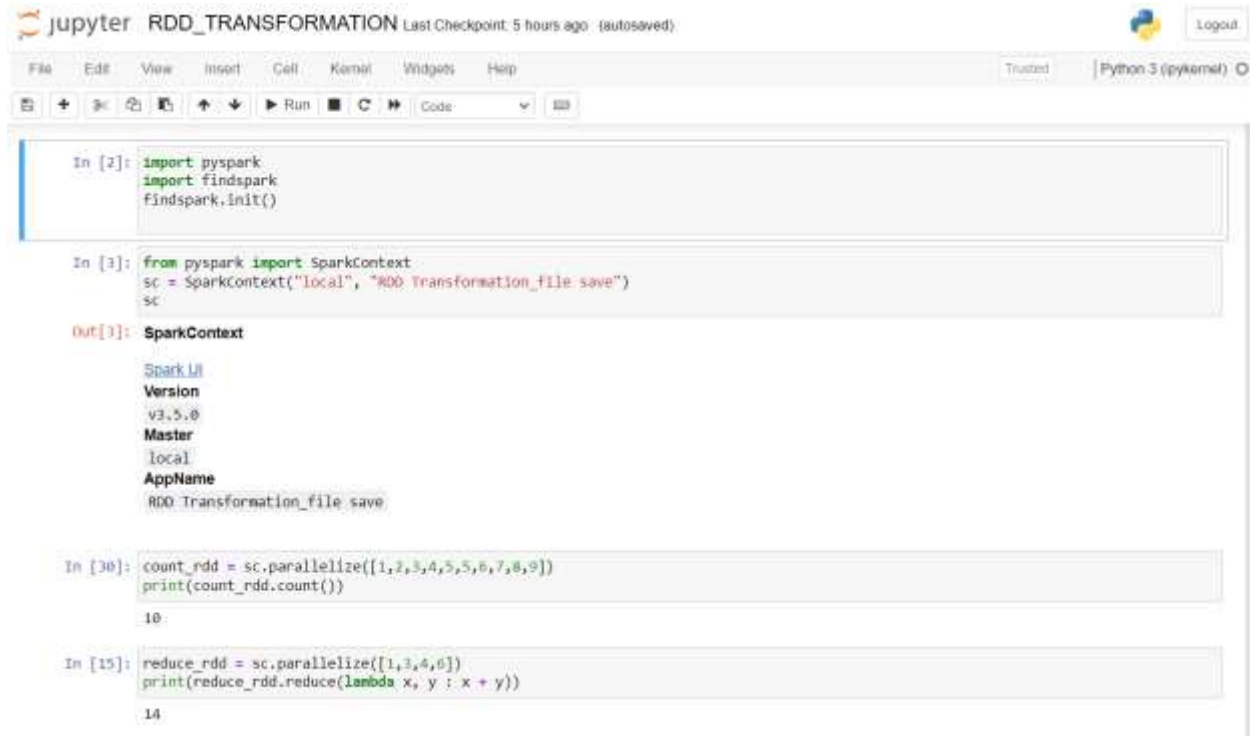
⑤ Add column: <var>.withColumn('col', expr())

C: Expr, (var('Expression'))

⑥ Drop column: <var>.drop('col')

⑦ Rename column: <var>.withColumnRenamed('col', 'new')

RDD Transformation:



Jupyter Notebook interface showing RDD initialization and basic transformations. The notebook is titled "RDD_TRANSFORMATION" and shows the following code cells:

```
In [2]: import pyspark
import findspark
findspark.init()

In [3]: from pyspark import SparkContext
sc = SparkContext("local", "RDD Transformation_file save")
sc

Out[3]: SparkContext

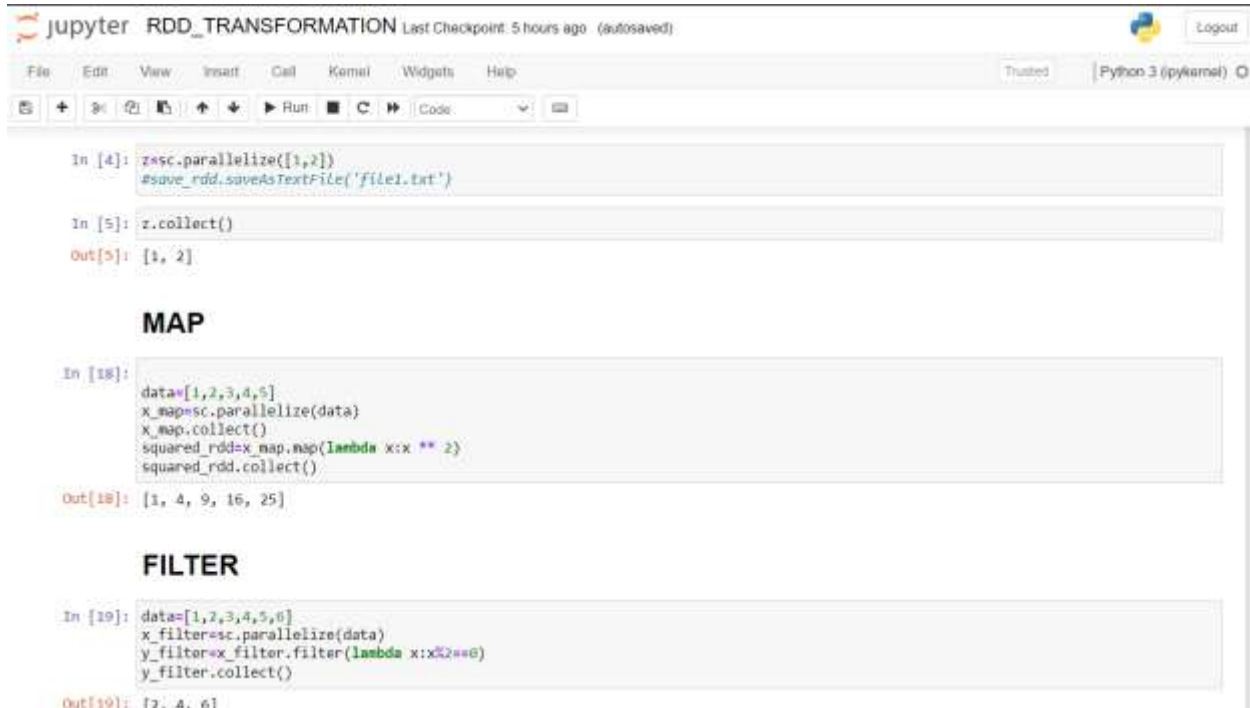
Spark UI
Version
v3.5.0
Master
local
AppName
RDD Transformation_file save

In [30]: count_rdd = sc.parallelize([1,2,3,4,5,5,6,7,8,9])
print(count_rdd.count())

10

In [15]: reduce_rdd = sc.parallelize([1,3,4,6])
print(reduce_rdd.reduce(lambda x, y : x + y))

14
```



Jupyter Notebook interface showing RDD transformations: MAP and FILTER. The notebook is titled "RDD_TRANSFORMATION" and shows the following code cells:

```
In [4]: z=sc.parallelize([1,2])
#save_rdd.saveAsTextFile('file1.txt')

In [5]: z.collect()

Out[5]: [1, 2]

MAP

In [18]: data=[1,2,3,4,5]
x_map=sc.parallelize(data)
x_map.collect()
squared_rdd=x_map.map(lambda x:x ** 2)
squared_rdd.collect()

Out[18]: [1, 4, 9, 16, 25]

FILTER

In [19]: data=[1,2,3,4,5,6]
x_filter=sc.parallelize(data)
y_filter=x_filter.filter(lambda x:x%2==0)
y_filter.collect()

Out[19]: [2, 4, 6]
```

flatMap

```
In [20]: data=[1,2,3,4,5]
x_flatmap=sc.parallelize(data)
y_flatmap=x_flatmap.flatMap(lambda x:(x,x**2,x**3))
y_flatmap.collect()

Out[20]: [1, 1, 1, 2, 2, 4, 8, 3, 9, 27, 4, 16, 64, 5, 25, 125]
```

Union ,Intersect,distinct

```
In [21]: data1=sc.parallelize([1,2,3,3,4,4,5])
data2=sc.parallelize([2,4,5])
union_rdd=data1.union(data2)
union_rdd.collect()

Out[21]: [1, 2, 3, 3, 4, 4, 5, 2, 4, 5]

In [22]: intersect_rdd=data1.intersection(data2)
intersect_rdd.collect()

Out[22]: [2, 4, 5]

In [23]: distinct_rdd=data1.distinct()
distinct_rdd.collect()

Out[23]: [1, 2, 3, 4, 5]
```

groupByKey,reduceByKey,sortByKey

```
In [24]: rdd = sc.parallelize([(1, 'apple'), (2, 'banana'), (3, 'orange')])
grouped=rdd.groupByKey()
grouped.collect()

Out[24]: [(1, <pyspark.resultiterable.ResultIterable at 0x23b8f2d4cd0>),
(2, <pyspark.resultiterable.ResultIterable at 0x23b9184e010>)]

In [25]: rdd = sc.parallelize([(1, 2), (2, 3), (3, 4),(2,5)])
reduced_rdd = rdd.reduceByKey(lambda x, y: x + y)
print("ReduceByKey RDD:", reduced_rdd.collect())

ReduceByKey RDD: [(1, 6), (2, 8)]

In [26]: rdd = sc.parallelize([(3, 'apple'), (1, 'banana'), (2, 'orange')])
sorted_rdd = rdd.sortByKey()
print("Sorted RDD:", sorted_rdd.collect())

Sorted RDD: [(1, 'banana'), (2, 'orange'), (3, 'apple')]
```

Sorted RDD: [(1, 'banana'), (2, 'orange'), (3, 'apple')]

mapvalues

```
In [27]: rdd = sc.parallelize([(1, 'apple'), (2, 'banana'), (3, 'orange')])
mapped_values_rdd = rdd.mapValues(lambda x: x.upper())
print("MapValues RDD:", mapped_values_rdd.collect())
```

MapValues RDD: [(1, 'APPLE'), (2, 'BANANA'), (3, 'ORANGE')]

```
In [28]: rdd = sc.parallelize([(1, 'apple'), (2, 'banana banana'), (3, 'orange')])
flat_mapped_values_rdd = rdd.flatMapValues(lambda x: x.split())
print("FlatMapValues RDD:", flat_mapped_values_rdd.collect())
```

FlatMapValues RDD: [(1, 'apple'), (2, 'banana'), (2, 'banana'), (3, 'orange')]

```
In [29]: sc1.stop()
```

```
jupyter DataFrame Last Checkpoint: 11 hours ago (autosaved) Logout
```

```
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)
```

```
In [38]: import os  
import sys  
  
os.environ['PYSPARK_PYTHON'] = sys.executable  
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable  
  
In [1]: import pyspark  
import findspark  
findspark.init()  
  
In [2]: from pyspark.sql import SparkSession  
  
In [3]: spark=SparkSession.builder.appName('DataFrame').getOrCreate()  
  
In [4]: spark  
Out[4]: SparkSession - in-memory  
SparkContext  
  
Spark UI  
Version  
v3.5.0  
Master  
local[*]  
AppName  
DataFrame
```

```
jupyter DataFrame Last Checkpoint: 11 hours ago (autosaved)
```

File Edit View Insert Cell Kernel Widgets Help

Not Trashed Python 3 (ipykernel) C

+ -> Run Code

```
In [35]: spark.read.csv("D:\Hexaware\Data_Engineering\Python\output_file.csv").show()
```

_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10	_c11
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	NULL	S
2	1	1	Cummings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	1	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2, 3101282	7.925	NULL	S
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	NULL	S
6	0	3	Moran, Mr. James	male	NULL	0	0	330877	8.4583	NULL	Q
7	0	1	McCarthy, Mr. Tin...	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	NULL	S
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	NULL	S
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	NULL	C
11	1	1	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	NULL	S
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	NULL	S
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	NULL	S
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	NULL	S
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	NULL	Q
18	1	2	Williams, Mr. Cha...	male	NULL	0	0	244373	13.0	NULL	S
19	0	3	Vander Planke, Mr...	female	31.0	1	0	385763	18.0	NULL	S

only showing top 20 rows

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3 (ipykernel)

+ - Run Stop Code

In [36]: df_pyspark=spark.read.option("header","true").csv('D:\Hexaware\Data_Engineering\Python\output_file.csv',inferSchema=True)

In [37]: df_pyspark.show()

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	NULL	S
2	1	1	Cummings, Mrs. Joh...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	NULL	S
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	NULL	S
6	0	3	Moran, Mr. James	male	NULL	0	0	330877	8.4583	NULL	Q
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	NULL	S
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	NULL	S
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	NULL	C
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	S
13	0	3	Saunderscock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	NULL	S
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	NULL	S
15	0	3	Vestrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	NULL	S
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	NULL	S
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	NULL	Q
18	1	2	Williams, Mr. Cha...	male	NULL	0	0	244373	13.0	NULL	S
19	0	3	Vander Planke, Mr...	female	31.0	1	0	345763	18.0	NULL	S
20	1	3	Masellmani, Mrs. ...	female	NULL	0	0	2649	7.225	NULL	C

only showing top 20 rows

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3 (ipykernel)

+ - Run Stop Code

```
In [38]: type(df_pyspark)
df_pyspark.printSchema()

root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

In [39]: df_pyspark=spark.read.csv('D:\Hexaware\Data_Engineering\Python\output_file.csv',header=True,inferSchema=True)

In [40]: #df_pyspark.show()
df_pyspark.printSchema()

```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
```


jupyter DataFrame Last Checkpoint: 11 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (ipykernel)

Run

Code

In [41]:

#select particular column
df_pyspark.select('Name').show()
df_pyspark.select(['Name', 'Age']).show()

+-----+
| Name|
+-----+
|Braund, Mr. Owen ...|
|Cumings, Mrs. Joh...|
|Heikkinen, Miss. ...|
|Futrelle, Mrs. Ja...|
|Allen, Mr. Willia...|
|Moran, Mr. James|
|McCarthy, Mr. Tim...|
|Palsson, Master. ...|
|Johnson, Mrs. Osc...|
|Nasser, Mrs. Rich...|
|Sandstrom, Miss. ...|
|Bonnell, Miss. El...|
|Saunderscock, Mr. ...|
|Andersson, Mr. An...|
|Vestrom, Miss. Hu...|
|Hewlett, Mrs. (Ma...|
|Rice, Master. Eugene|
|Williams, Mr. Cha...|
|Vander Planke, Mr...|
|Masselmani, Mrs. ...|
+-----+
only showing top 20 rows

jupyter DataFrame Last Checkpoint: 11 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (ipykernel)

Run

Code

In [42]:

#check datatypes
df_pyspark.dtypes

Out[42]:

[('PassengerId', 'int'),
('Survived', 'int'),
('Pclass', 'int'),
('Name', 'string'),
('Sex', 'string'),
('Age', 'double'),
('SibSp', 'int'),
('Parch', 'int'),
('Ticket', 'string'),
('Fare', 'double'),
('Cabin', 'string'),
('Embarked', 'string')]

In [43]:

#descr(bv
df_pyspark.describe().show()

+-----+
|summary| PassengerId| Survived| Pclass| Name| Sex| Age| S
|ibSp| Parch| Ticket| Fare|Cabin|Embarked|
+-----+
count	891	891	891	891	891	714
mean	891	891	891	891	204	889
446.0	0.3838383838383838	2.308641075308642				
1896	0.38159371492704824	260318.54016792738	32.2042079685746	NULL	NULL	
stddev	257.3538420152301	0.48659245426485753	0.8360712409770491	NULL	NULL	14.526497332334035
4315	0.8060572211295488	471609.26868834975	49.69342859718089	NULL	NULL	
min	1	0	1	"Andersson, Mr. A...	female	0.42

In [44]: df_pyspark.withColumn('Age*2',df_pyspark['Age']*2).show()

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Age*2
1	0	3	Braund, Mr. Owen ...	male	22.0	1	0	A/5 21171	7.25	NULL	S	44.0
2	1	1	Cumings, Mrs. Joh...	female	38.0	1	0	PC 17590	71.2833	C85	C	76.0
3	1	3	Heikkinen, Miss. ...	female	26.0	0	0	STON/O2. 3101282	7.925	NULL	S	52.0
4	1	1	Futrelle, Mrs. Ja...	female	35.0	1	0	113803	53.1	C123	S	70.0
5	0	3	Allen, Mr. Willia...	male	35.0	0	0	373450	8.05	NULL	S	70.0
6	0	3	Moran, Mr. James	male	NULL	0	0	338877	8.4583	NULL	Q	NULL
7	0	1	McCarthy, Mr. Tim...	male	54.0	0	0	17463	51.8625	E66	S	108.0
8	0	3	Palsson, Master. ...	male	2.0	3	1	349909	21.075	NULL	S	4.0
9	1	3	Johnson, Mrs. Osc...	female	27.0	0	2	347742	11.1333	NULL	S	54.0
10	1	2	Nasser, Mrs. Nich...	female	14.0	1	0	237736	30.0708	NULL	C	28.0
11	1	3	Sandstrom, Miss. ...	female	4.0	1	1	PP 9540	16.7	G6	S	8.0
12	1	1	Bonnell, Miss. El...	female	58.0	0	0	113783	26.55	C103	S	116.0
13	0	3	Saunderscock, Mr. ...	male	20.0	0	0	A/5. 2151	8.05	NULL	S	40.0
14	0	3	Andersson, Mr. An...	male	39.0	1	5	347082	31.275	NULL	S	78.0
15	0	3	Vostrom, Miss. Hu...	female	14.0	0	0	350406	7.8542	NULL	S	28.0
16	1	2	Hewlett, Mrs. (Ma...	female	55.0	0	0	248706	16.0	NULL	S	110.0
17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125	NULL	Q	4.0
18	1	2	Williams, Mr. Cha...	male	NULL	0	0	244373	13.0	NULL	S	NULL
19	0	3	Vander Planke, Mr...	female	31.0	1	0	345763	18.0	NULL	S	62.0
20	1	3	Maslemanni, Mrs. ...	female	NULL	0	0	2649	7.225	NULL	C	NULL

only showing top 20 rows

In [45]: df_pyspark.drop('Age*2')

Out[45]: DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Sex: string, Age: double, SibSp: int, Parch: int, Ticket: string, Fare: double, Cabin: string, Embarked: string]

In [46]: df_pyspark.withColumnRenamed('Sex','Gender')

Out[46]: DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Gender: string, Age: double, SibSp: int, Parch: int, Ticket: string, Fare: double, Cabin: string, Embarked: string]

In [5]:

```
# Create data in dataframe
data = [
    (('Ram', '1991-04-01', 'M', 3000),
     ('Mike', '2000-05-19', 'M', 4000),
     ('Rohini', '1978-09-05', 'M', 4000),
     ('Maria', '1967-12-01', 'F', 4000),
     ('Jenis', '1980-02-17', 'F', 1200))

# Column names in dataframe
columns = ["Name", "DOB", "Gender", "salary"]

# Create the spark dataframe
df = spark.createDataFrame(data=data,
                           schema=columns)

# Print the dataframe
df.show()
```

Name	DOB	Gender	salary
Ram	1991-04-01	M	3000
Mike	2000-05-19	M	4000
Rohini	1978-09-05	M	4000
Maria	1967-12-01	F	4000
Jenis	1980-02-17	F	1200

In [20]: spark.stop()