NAME: MAVIA ALAM KHAN(2303.KHI.DEG.017)

PAIRING WITH **: Mohammad Hussam (2303.KHI.DEG.020)**

ASSIGNMENT NO : 5.3

Read data from source to DataFrame in local Spark setup and display
DataFrame schema.
tasks/5_data_pipelines/day_3_spark/data_assignment
For numerical columns, calculate minimum, maximum and average values.
For categorical columns, create and apply UDF that will change the last letter of
every word to “1”.
Sort DataFrame by the first column and save the results to the Parquet file.

SOLUTION:

STEP 1 :

Importing the necessary modules / libraries .

Import the necessary modules

[1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *
```

Creating spark session

creating spark session

```
spark = SparkSession.builder.appName("PySpark_assign").getOrCreate()
```

```
23/05/21 17:35:52 WARN Utils: Your hostname, mhussam resolves to a loopback address: 127.0.1.1; using 192.168.185.62 instead (on interface wlp3s0)
23/05/21 17:35:52 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/05/21 17:35:53 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/05/21 17:35:54 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/05/21 17:35:54 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
23/05/21 17:35:54 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
23/05/21 17:35:54 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
23/05/21 17:35:54 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
23/05/21 17:35:54 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
```

STEP 2 :

Reading the data from the CSV file into a DataFrame.

Reading the data from the CSV file into a DataFrame

```
titanic = spark.read.option("header", "true").option("inferSchema", "true").option("header", "false").csv("data/titanic.csv")
## Defining  the column names for header row
columns = ["PassengerId", "Survived", "Pclass", "Name", "Sex", "Age", "SibSp", "Parch", "Ticket", "Fare", "Cabin", "Embarked", "Timestamp"
#Adding the column names as header row in the DataFrame
titanic = titanic.toDF(*columns)
titanic.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Timestamp|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|          1|       0|     3|Braund, Mr. Owen ...|  male|  22|    1|    0|       A/5 21171|   7.25| null|       S|2020-01-01 13:45:25|
|          2|       1|     1|Cumings, Mrs. Joh...|female|  38|    1|    0|        PC 17599|71.2833|  C85|       C|2020-01-01 13:44:48|
|          3|       1|     3|Heikkinen, Miss. ...|female|  26|    0|    0|STON/O2. 3101282|  7.925| null|       S|2020-01-01 13:38:11|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|  35|    1|    0|          113803|   53.1| C123|       S|2020-01-01 13:32:00|
|          5|       0|     3|Allen, Mr. Willia...|  male|  35|    0|    0|          373450|   8.05| null|       S|2020-01-01 13:36:30|
|          6|       0|     3|    Moran, Mr. James|  male|null|    0|    0|          330877| 8.4583| null|       Q|2020-01-01 13:31:39|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|  54|    0|    0|           17463|51.8625|  E46|       S|2020-01-01 13:37:31|
|          8|       0|     3|Palsson, Master. ...|  male|   2|    3|    1|          349909| 21.075| null|       S|2020-01-01 13:49:08|
|          9|       1|     3|Johnson, Mrs. Osc...|female|  27|    0|    2|          347742|11.1333| null|       S|2020-01-01 13:33:42|
|         10|       1|     2|Nasser, Mrs. Nich...|female|  14|    1|    0|          237736|30.0708| null|       C|2020-01-01 13:32:53|
|         11|       1|     3|Sandstrom, Miss. ...|female|   4|    1|    1|         PP 9549|   16.7|   G6|       S|2020-01-01 13:32:23|
|         12|       1|     1|Bonnell, Miss. El...|female|  58|    0|    0|          113783|  26.55| C103|       S|2020-01-01 13:30:12|
|         13|       0|     3|Saundercock, Mr. ...|  male|  20|    0|    0|       A/5. 2151|   8.05| null|       S|2020-01-01 13:33:34|
|         14|       0|     3|Andersson, Mr. An...|  male|  39|    1|    5|          347082| 31.275| null|       S|2020-01-01 13:30:20|
|         15|       0|     3|Vestrom, Miss. Hu...|female|  14|    0|    0|          350406| 7.8542| null|       S|2020-01-01 13:41:17|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|  55|    0|    0|          248706|   16.0| null|       S|2020-01-01 13:34:22|
|         17|       0|     3|Rice, Master. Eugene|  male|   2|    4|    1|          382652| 29.125| null|       Q|2020-01-01 13:41:55|
|         18|       1|     2|Williams, Mr. Cha...|  male|null|    0|    0|          244373|   13.0| null|       S|2020-01-01 13:39:35|
|         19|       0|     3|Vander Planke, Mr...|female|  31|    1|    0|          345763|   18.0| null|       S|2020-01-01 13:39:38|
|         20|       1|     3|Masselmani, Mrs. ...|female|null|    0|    0|            2649|  7.225| null|       C|2020-01-01 13:36:56|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
only showing top 20 rows
```

STEP 3 :

After that we will display the Dataframe schema

only showing top 20 rows

Displaying the Dataframe schema

```
[13]: titanic.printSchema()
```

```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: integer (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)
 |-- Timestamp: timestamp (nullable = true)
```

STEP 4 :

Changing the data type of the "Survived" column to string and converting the binary values into strings for categorization purposes.

modified the schema by changing the data type of the "Survived" column to string and converting the binary values into strings for categorization purposes ¶

```
titanic = titanic.withColumn("Survived", when(col("Survived") == 1, "YES").otherwise("NO").cast("string"))
titanic.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Timestamp|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|          1|      NO|     3|Braund, Mr. Owen ...|  male|  22|    1|    0|       A/5 21171|   7.25| null|       S|2020-01-01 13:45:25|
|          2|     YES|     1|Cumings, Mrs. Joh...|female|  38|    1|    0|        PC 17599|71.2833|  C85|       C|2020-01-01 13:44:48|
|          3|     YES|     3|Heikkinen, Miss. ...|female|  26|    0|    0|STON/O2. 3101282|  7.925| null|       S|2020-01-01 13:38:11|
|          4|     YES|     1|Futrelle, Mrs. Ja...|female|  35|    1|    0|          113803|   53.1| C123|       S|2020-01-01 13:32:00|
|          5|      NO|     3|Allen, Mr. Willia...|  male|  35|    0|    0|          373450|   8.05| null|       S|2020-01-01 13:36:30|
|          6|      NO|     3|    Moran, Mr. James|  male|null|    0|    0|          330877| 8.4583| null|       Q|2020-01-01 13:31:39|
|          7|      NO|     1|McCarthy, Mr. Tim...|  male|  54|    0|    0|           17463|51.8625|  E46|       S|2020-01-01 13:37:31|
|          8|      NO|     3|Palsson, Master. ...|  male|   2|    3|    1|          349909| 21.075| null|       S|2020-01-01 13:49:08|
|          9|     YES|     3|Johnson, Mrs. Osc...|female|  27|    0|    2|          347742|11.1333| null|       S|2020-01-01 13:33:42|
|         10|     YES|     2|Nasser, Mrs. Nich...|female|  14|    1|    0|          237736|30.0708| null|       C|2020-01-01 13:32:53|
|         11|     YES|     3|Sandstrom, Miss. ...|female|   4|    1|    1|         PP 9549|   16.7|   G6|       S|2020-01-01 13:32:23|
|         12|     YES|     1|Bonnell, Miss. El...|female|  58|    0|    0|          113783|  26.55| C103|       S|2020-01-01 13:30:12|
|         13|      NO|     3|Saundercock, Mr. ...|  male|  20|    0|    0|       A/5. 2151|   8.05| null|       S|2020-01-01 13:33:34|
|         14|      NO|     3|Andersson, Mr. An...|  male|  39|    1|    5|          347082| 31.275| null|       S|2020-01-01 13:30:20|
|         15|      NO|     3|Vestrom, Miss. Hu...|female|  14|    0|    0|          350406| 7.8542| null|       S|2020-01-01 13:41:17|
|         16|     YES|     2|Hewlett, Mrs. (Ma...|female|  55|    0|    0|          248706|   16.0| null|       S|2020-01-01 13:34:22|
|         17|      NO|     3|Rice, Master. Eugene|  male|   2|    4|    1|          382652| 29.125| null|       Q|2020-01-01 13:41:55|
|         18|     YES|     2|Williams, Mr. Cha...|  male|null|    0|    0|          244373|   13.0| null|       S|2020-01-01 13:39:35|
|         19|      NO|     3|Vander Planke, Mr...|female|  31|    1|    0|          345763|   18.0| null|       S|2020-01-01 13:39:38|
|         20|     YES|     3|Masselmani, Mrs. ...|female|null|    0|    0|            2649|  7.225| null|       C|2020-01-01 13:36:56|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
only showing top 20 rows
```

Now the updated schema after above :

6]:
```
## updated schema
titanic.printSchema()
```

```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: string (nullable = false)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: integer (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)
 |-- Timestamp: timestamp (nullable = true)
```

STEP 5 :

Calculating the minimum, maximum and mean values for the numerical data.

Calculating the minimum, maximum and mean values of the numerical data

[9]:
```
numerical_cols = [col_name for col_name, col_type in titanic.dtypes if col_type in ['int', 'bigint', 'float', 'double']]
numerical_df = titanic.select(*numerical_cols)


min_max_mean=numerical_df.describe()
min_max_mean.select(numerical_cols).summary('min','max','mean').show()
```

```
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|summary|       PassengerId|            Pclass|               Age|             SibSp|             Parch|              Fare|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|    min|                 1|0.8360712409770491|                 0|                 0|                 0|               0.0|
|    max|               891|               891|                80|               891|               891|               891|
|   mean|497.27076840304596|179.62894264325715|167.64315089562422|180.12515025772694|179.63753018721 14|297.04536731315113|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
```

STEP 6:

For categorical columns, we create and apply UDF that will change the last letter of every word to “1”

For categorical columns, create and apply UDF that will change the last letter of every word to “1”

[12]:
```
cat_columns = ["Sex","Cabin","Embarked","Survived"]

def change_last_letter_after_space(word):
    if word is not None:
        words = word.split()
        for i in range(len(words)):
            words[i] = words[i][:-1] + "1"
        return " ".join(words)
    return word
change_last_letter_udf = udf(change_last_letter_after_space, StringType())
for column in cat_columns:
    titanic = titanic.withColumn(column, change_last_letter_udf(titanic[column]))
titanic.show(10)
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Timestamp|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|          1|      N1|     3|Braund, Mr. Owen ...|  mal1|  22|    1|    0|       A/5 21171|   7.25| null|       1|2020-01-01 13:45:25|
|          2|     YE1|     1|Cumings, Mrs. Joh...|femal1|  38|    1|    0|        PC 17599|71.2833|  C81|       1|2020-01-01 13:44:48|
|          3|     YE1|     3|Heikkinen, Miss. ...|femal1|  26|    0|    0|STON/O2. 3101282|  7.925| null|       1|2020-01-01 13:38:11|
|          4|     YE1|     1|Futrelle, Mrs. Ja...|femal1|  35|    1|    0|          113803|   53.1| C121|       1|2020-01-01 13:32:00|
|          5|      N1|     3|Allen, Mr. Willia...|  mal1|  35|    0|    0|          373450|   8.05| null|       1|2020-01-01 13:36:30|
|          6|      N1|     3|    Moran, Mr. James|  mal1|null|    0|    0|          330877| 8.4583| null|       1|2020-01-01 13:31:39|
|          7|      N1|     1|McCarthy, Mr. Tim...|  mal1|  54|    0|    0|           17463|51.8625|  E41|       1|2020-01-01 13:37:31|
|          8|      N1|     3|Palsson, Master. ...|  mal1|   2|    3|    1|          349909| 21.075| null|       1|2020-01-01 13:49:08|
|          9|     YE1|     3|Johnson, Mrs. Osc...|femal1|  27|    0|    2|          347742|11.1333| null|       1|2020-01-01 13:33:42|
|         10|     YE1|     2|Nasser, Mrs. Nich...|femal1|  14|    1|    0|          237736|30.0708| null|       1|2020-01-01 13:32:53|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
only showing top 10 rows
```

STEP 7 :

sorting the df by first column and saving the results to the Parquet file.

sorting the df by first column

```
[15]: sorted_df = titanic.orderBy(col(titanic.columns[0]))
```

saving the results to the Parquet file.

```
[17]: sorted_df.write.mode("overwrite").parquet("output/titanic_data.parquet")
```

Saved parquet in directory as :