

Name: - Muhammad Huzaifa Waseem (2303-KHI-DEG-021)

Pair Partner 1: - Muhammad Faizan Rafique (2303.005.KHI.DEG)

Pair Partner 2: - Syed Muhammad Hammad Irshad(2303.KHI.DEG.032)

UNIT 5.3:

Assignment

Read data from source to DataFrame in local Spark setup and display DataFrame schema.

tasks/4_data_pipelines/day_3_spark/data_assignment

The screenshot shows a Jupyter Notebook with the following code and output:

```
[55]: import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, mean, max, min, udf
from pyspark.sql.types import DateType, IntegerType, StringType, StructType, FloatType

[56]: spark = SparkSession.builder \
    .appName("DataAssignment") \
    .getOrCreate()

[57]: df = spark.read.csv("titanic.csv", header=False, inferSchema=True)

[58]: df.show()
```

	_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10	_c11	_c12
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	Cunings, Mrs. Joh...	female	38	1	0		PC 17599	71.2833	C85	C	2020-01-01 13:44:48
3	1	1	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		373458	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. Tin...	male	54	0	0		17463	51.8625	E46	S	2020-01-01 13:37:31
8	0	3	Palsson, Master. ...	male	21	3	1		349909	21.075	null	S	2020-01-01 13:49:00
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	41	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	Saunderscock, 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		244573	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	MasseMani, Mrs. ...	female	null	0	0		2649	7.225	null	C	2020-01-01 13:36:56

only showing top 20 rows

```
[59]: df.printSchema()

root
|-- _c0: integer (nullable = true)
|-- _c1: integer (nullable = true)
|-- _c2: integer (nullable = true)
|-- _c3: string (nullable = true)
|-- _c4: string (nullable = true)
|-- _c5: integer (nullable = true)
|-- _c6: integer (nullable = true)
|-- _c7: integer (nullable = true)
|-- _c8: string (nullable = true)
|-- _c9: double (nullable = true)
|-- _c10: string (nullable = true)
|-- _c11: string (nullable = true)
|-- _c12: timestamp (nullable = true)
```

For numerical columns, calculate minimum, maximum and average values.

```
127.0.0.1:8888/lab/tree/workspace/ok.ipynb

File Edit View Run Kernel Tabs Settings Help

+ ok.ipynb
- workspace /
  Name Last Modified
  output.par... 2 minutes ago
  ok.ipynb a minute ago
  titanic.csv 2 days ago

[60]: # Get the column names and data types
      column_types = df.dtypes
      # Filter the DataFrame to select only integer columns
      integer_columns = [col_name for col_name, col_type in column_types if col_type == "int" or col_type == "double"]
      # Select the integer columns
      selected_integer_columns_df = df.select(*integer_columns)
      # Show the selected DataFrame
      selected_integer_columns_df.show()

+-----+-----+-----+-----+-----+-----+
|_c0|_c1|_c2|_c5|_c6|_c7|_c9|
+-----+-----+-----+-----+-----+
| 1| 0| 3| 22| 1| 0| 7.25|
| 2| 1| 1| 38| 1| 0| 71.2833|
| 3| 1| 3| 26| 0| 0| 7.925|
| 4| 1| 1| 35| 1| 0| 53.1|
| 5| 0| 3| 35| 0| 0| 8.05|
| 6| 0| 3| null| 0| 0| 8.4583|
| 7| 0| 1| 54| 0| 0| 51.8625|
| 8| 0| 3| 2| 3| 1| 21.075|
| 9| 1| 3| 27| 0| 2| 11.1333|
|10| 1| 2| 14| 1| 0| 30.0708|
|11| 1| 3| 4| 1| 1| 16.7|
|12| 1| 1| 58| 0| 0| 26.55|
|13| 0| 3| 20| 0| 0| 8.05|
|14| 0| 3| 39| 1| 5| 31.275|
|15| 0| 3| 14| 0| 0| 7.8542|
|16| 1| 2| 55| 0| 0| 16.0|
|17| 0| 3| 2| 4| 1| 29.125|
|18| 1| 2| null| 0| 0| 13.0|
|19| 0| 3| 31| 1| 0| 18.0|
|20| 1| 3| null| 0| 0| 7.225|
+-----+-----+-----+-----+-----+
only showing top 20 rows

[61]: initial_state = True
      for col in selected_integer_columns_df:
          calculate_df = df.agg(mean(col).alias("Average"), max(col).alias("Max"), min(col).alias("Min"))
          if initial_state:
              merged_avg_max_min_df = calculate_df
              initial_state = False
          else:
              merged_avg_max_min_df = merged_avg_max_min_df.unionAll(calculate_df)
      merged_df.show()

+-----+-----+-----+-----+-----+
| Average| Max| Min|
+-----+-----+-----+-----+
| 446.0| 891.0| 1.0|
| 0.3838383838383838| 1.0| 0.0|
| 2.388641975388642| 3.0| 1.0|
| 29.67927188683473| 80.0| 0.0|
| 0.5238078563411896| 8.0| 0.0|
| 0.38159371492704824| 6.0| 0.0|
| 32.2042079685746| 512.3292| 0.0|
+-----+-----+-----+-----+-----+
```

For categorical columns, create and apply UDF that will change the last letter of every word to "I".

```
127.0.0.1:8888/lab/tree/workspace/ok.ipynb

File Edit View Run Kernel Tabs Settings Help

+ ok.ipynb
- workspace /
  Name Last Modified
  output.par... 3 minutes ago
  ok.ipynb 2 minutes ago
  titanic.csv 2 days ago

[62]: str_columns = ["_c4", "_c10", "_c11"]
      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] + "I"
              return " ".join(words)
          return word
      change_last_letter_udf = udf(change_last_letter_after_space, StringType())
      for column in str_columns:
          df = df.withColumn(column, change_last_letter_udf(df[column]))
      df.show()

+-----+-----+-----+-----+-----+-----+-----+
|_c0|_c1|_c2|_c3|_c4|_c5|_c6|_c7|_c8|_c9|_c10|_c11|_c12|
+-----+-----+-----+-----+-----+-----+-----+
| 1| 0| 3| Braund, Mr. Owen ...| male| 22| 1| 0| A/5 21171| 7.25| null| 1| 2020-01-01 13:45:25|
| 2| 1| 1| Cumings, Mrs. Joh...| female| 38| 1| 0| PC 17599| 71.2833| C81| 1| 2020-01-01 13:44:48|
| 3| 1| 3| Heikkinen, Miss. ...| female| 26| 0| 0| STON/O2. 3101282| 7.925| null| 1| 2020-01-01 13:38:11|
| 4| 1| 1| Futrelle, Mrs. Ja...| female| 35| 1| 0| 113803| 53.1| C121| 1| 2020-01-01 13:32:00|
| 5| 0| 3| Allen, Mr. Willia...| male| 35| 0| 0| 373450| 8.05| null| 1| 2020-01-01 13:36:30|
| 6| 0| 3| Moran, Mr. James...| male| null| 0| 0| 330877| 8.4583| null| 1| 2020-01-01 13:31:39|
| 7| 0| 1| McCarthy, Mr. Tim...| male| 54| 0| 0| 17463| 51.8625| E41| 1| 2020-01-01 13:37:31|
| 8| 0| 3| Palsson, Master. ...| male| 2| 3| 1| 349909| 21.075| null| 1| 2020-01-01 13:49:08|
| 9| 1| 3| Johnson, Mrs. Osc...| female| 27| 0| 2| 347742| 11.1333| null| 1| 2020-01-01 13:33:42|
|10| 1| 2| Hassner, Mrs. Nich...| female| 14| 1| 0| 237736| 30.0708| null| 1| 2020-01-01 13:32:53|
|11| 1| 3| Sandstrom, Miss. ...| female| 4| 1| 1| PP 9549| 16.7| G1| 1| 2020-01-01 13:32:23|
|12| 1| 1| Bonnell, Miss. EL...| female| 58| 0| 0| 113783| 26.55| C101| 1| 2020-01-01 13:30:12|
|13| 0| 3| Saunderson, Mr. ...| male| 20| 0| 0| A/5. 2151| 8.05| null| 1| 2020-01-01 13:33:34|
|14| 0| 3| Andersson, Mr. An...| male| 39| 1| 5| 347082| 31.275| null| 1| 2020-01-01 13:30:20|
|15| 0| 3| Westrom, Miss. Hu...| female| 14| 0| 0| 350406| 7.8542| null| 1| 2020-01-01 13:41:17|
|16| 1| 2| Hewlett, Mrs. (Ma...| female| 55| 0| 0| 248706| 16.0| null| 1| 2020-01-01 13:34:22|
|17| 0| 3| Rice, Master. Eugene| male| 2| 4| 1| 382652| 29.125| null| 1| 2020-01-01 13:41:55|
|18| 1| 2| Williams, Mr. Cha...| male| null| 0| 0| 244373| 13.0| null| 1| 2020-01-01 13:39:35|
|19| 0| 3| Vander Planke, Mr...| female| 31| 1| 0| 345763| 18.0| null| 1| 2020-01-01 13:39:38|
|20| 1| 3| Masselmani, Mrs. ...| female| null| 0| 0| 2649| 7.225| null| 1| 2020-01-01 13:36:56|
+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

Sort DataFrame by the first column and save the results to the Parquet file.

127.0.0.1:8888/lab/tree/workspace/ok.ipynb

File Edit View Run Kernel Tabs Settings Help

Filter files by name

/ workspace /

Name	Last Modified
output.par...	3 minutes ago
ok.ipynb	2 minutes ago
titanic.csv	2 days ago

```
[63]: # Sort DataFrame by the first column
sorted_df = df.orderBy("_c0")
sorted_df.show()
```

_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10	_c11	_c12	
1	0	3	Braund, Mr. Owen ...	male	22	1	0	A/5	21171	7.25	null	1 2020-01-01 13:45:25	
2	1	1	Cunings, Mrs. Joh...	female	38	1	0	PC	17599	71.2833	C81	1 2020-01-01 13:44:48	
3	1	3	Heikkinen, Miss. ...	female	26	0	0	STON/O2.	3101282	7.925	null	1 2020-01-01 13:38:11	
4	1	1	Futrelle, Mrs. Ja...	female	35	1	0		113803	53.1	C121	1 2020-01-01 13:32:00	
5	0	3	Allen, Mr. Willia...	male	35	0	0		373450	8.05	null	1 2020-01-01 13:36:30	
6	0	3	Moran, Mr. James	male	null	0	0		330877	8.4583	null	1 2020-01-01 13:31:39	
7	0	1	McCarthy, Mr. Tim...	male	54	0	0		17463	51.8625	E41	1 2020-01-01 13:37:31	
8	0	3	Palsson, Master. ...	male	21	3	1		349909	21.075	null	1 2020-01-01 13:49:08	
9	1	3	Johnson, Mrs. Osc...	female	27	0	2		347742	11.1333	null	1 2020-01-01 13:33:42	
10	1	2	Nasser, Mrs. Nich...	female	14	1	0		237736	30.0708	null	1 2020-01-01 13:32:53	
11	1	3	Sandstrom, Miss. ...	female	4	1	1		PP	9549	16.7	61	1 2020-01-01 13:32:23
12	1	1	Bonnell, Miss. El...	female	58	0	0		1137831	26.55	C101	1 2020-01-01 13:30:12	
13	0	3	Saunderscock, Mr. ...	male	20	0	0	A/5.	2151	8.05	null	1 2020-01-01 13:33:34	
14	0	3	Andersson, Mr. An...	male	39	1	5		347082	31.275	null	1 2020-01-01 13:30:20	
15	0	3	Vestrom, Miss. Hu...	female	14	0	0		350406	7.8542	null	1 2020-01-01 13:41:17	
16	1	2	Hewlett, Mrs. (Ma...	female	55	0	0		248786	16.0	null	1 2020-01-01 13:34:22	
17	0	3	Rice, Master. Eugene	male	2	4	1		382652	29.125	null	1 2020-01-01 13:41:55	
18	1	2	Williams, Mr. Cha...	male	null	0	0		244373	13.0	null	1 2020-01-01 13:39:35	
19	0	3	Vander Planke, Mr...	female	31	1	0		345763	18.0	null	1 2020-01-01 13:39:38	
20	1	3	Masselmani, Mrs. ...	female	null	0	0		2649	7.225	null	1 2020-01-01 13:36:56	

only showing top 20 rows

```
[65]: # Save sorted DataFrame to Parquet file
output_path = "output.parquet"
sorted_df.write.parquet(output_path)
```

output.parquet 4 minutes ago

ok.ipynb 3 minutes ago

titanic.csv 2 days ago

_SUCCESS 4 minutes ago

part-00000-f28a0229-cb60-4ec0-ba83-579455498006-c000.snappy.parquet 4 minutes ago