Welcome to exercise three of “Apache Spark for Scalable Machine Learning on BigData”. In this exercise you’ll create a DataFrame, register a temporary query table and issue SQL commands against it.

Let’s create a little data frame:

In [ ]:
```
from pyspark.sql import Row

df = spark.createDataFrame([Row(id=1, value='value1'),Row(id=2, value=
'value2')])

# let's have a look what's inside
df.show()

# let's print the schema
df.printSchema()
```

Now we register this DataFrame as query table and issue an SQL statement against it. Please note that the result of the SQL execution returns a new DataFrame we can work with.

In [ ]:
```
# register dataframe as query table
df.createOrReplaceTempView('df_view')

# execute SQL query
df_result = spark.sql('select value from df_view where id=2')

# examine contents of result
df_result.show()

# get result as string
df_result.first().value
```

Although we’ll learn more about DataFrames next week, please try to find a way to count the rows in this DataFrame by looking at the API documentation. No worries, we’ll cover DataFrames in more detail next week.

https://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.DataFrame

In [ ]:
```
df.$$$
```