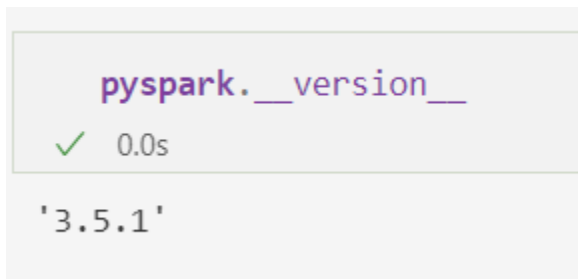


## Question 1:

Install Spark and PySpark

- Install Spark
- Run PySpark
- Create a local spark session
- Execute `spark.version`.

What's the output?



The screenshot shows a Jupyter Notebook cell with the following content:

```
pyspark.__version__
```

Below the code, there is a green checkmark and the text "0.0s", indicating successful execution. The output of the cell is displayed below a horizontal line:

```
'3.5.1'
```

## Question 2:

FHV October 2019

Read the October 2019 FHV into a Spark Dataframe with a schema as we did in the lessons.

Repartition the Dataframe to 6 partitions and save it to parquet.

What is the average size of the Parquet (ending with .parquet extension) Files that were created (in MB)? Select the answer which most closely matches.

- **1MB**
- 6MB
- 25MB
- 87MB

### Question 3:

Count records

How many taxi trips were there on the 15th of October?

Consider only trips that started on the 15th of October.

- 108,164
- 12,856
- 452,470
- **62,610**

Important

Be aware of columns order when defining schema

```
from pyspark.sql import functions as F
df \
    .withColumn('pickup_date', F.to_date(df.pickup_datetime)) \
    .withColumn('dropoff_date', F.to_date(df.dropoff_datetime)) \
    .select('pickup_date', 'dropoff_date', 'PUlocationID', 'DOlocationID') \
    .filter(F.to_date(df.pickup_datetime) == "2019-10-15") \
    .count()
```

✓ 3.3s

62610

### Question 4:

Longest trip for each day

What is the length of the longest trip in the dataset in hours?

- **631,152.50 Hours**
- 243.44 Hours
- 7.68 Hours
- 3.32 Hours

```
df \
  .withColumn('duration', ((df.dropoff_datetime.cast('long') - df.pickup_datetime.cast('long'))/3600)) \
  .withColumn('pickup_date', F.to_date(df.pickup_datetime)) \
  .groupBy('pickup_date') \
  | .max('duration') |
  .orderBy('max(duration)', ascending=False) \
  .limit(1) \
  .show()
```

✓ 5.2s

pickup_date	max(duration)
2019-10-11	631152.5

## Question 5:

User Interface

Spark's User Interface which shows the application's dashboard runs on which local port?

- 80
- 443
- **4040**
- 8080



## Question 6:

Least frequent pickup location zone

Load the zone lookup data into a temp view in Spark

### [Zone Data](#)

Using the zone lookup data and the FHV October 2019 data, what is the name of the LEAST frequent pickup location Zone?

- East Chelsea
- **Jamaica Bay**
- Union Sq
- Crown Heights North

```
zpu = df_zones \
    .withColumnRenamed('Zone', 'PUzone') \
    .withColumnRenamed('LocationID', 'zPULocationID') \
    .withColumnRenamed('Borough', 'PUBorough') \
    .drop('service_zone')
zdo = df_zones \
    .withColumnRenamed('Zone', 'DOzone') \
    .withColumnRenamed('LocationID', 'zDOLocationID') \
    .withColumnRenamed('Borough', 'DOBorough') \
    .drop('service_zone')

df_join_temp = df.join(zpu, df.PULocationID == zpu.zPULocationID)
df_join = df_join_temp.join(zdo, df_join_temp.DOLocationID == zdo.zDOLocationID)
```

✓ 0.2s

```
dd = df_join.drop('PULocationID', 'DOLocationID', 'zPULocationID', 'zDOLocationID')
```

✓ 0.0s

```
dd.createOrReplaceTempView ('join_table')
```

✓ 0.0s

```
spark.sql("""
```

```
SELECT
```

```
  PUzone,  
  COUNT(1)
```

```
FROM
```

```
  join_table
```

```
GROUP BY
```

```
  1
```

```
ORDER BY
```

```
  2 ASC
```

```
LIMIT
```

```
  1
```

```
;
```

```
""").show()
```

✓ 10.6s

```
+-----+-----+  
|      PUzone|count(1)|  
+-----+-----+  
|Jamaica Bay|        1|  
+-----+-----+
```