PKA-Q1-Q4

January 11, 2024

[1]:
```
import pyspark
import os
import sys
from pyspark import SparkContext
from pyspark import SparkConf
os.environ['PYSPARK_PYTHON'] = sys.executable
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable
```

[2]:
```
from pyspark.sql import SparkSession

# Create a SparkSession
spark = SparkSession.builder.getOrCreate()

# Create a DataFrame
df = spark.createDataFrame([
    (1, "Alice"),
    (2, "Bob"),
    (3, "Carol"),
    (4, "Dave"),
    (5, "Eve")
], ["id", "name"])
print(df.count())
df.show()
#df.write.format("csv").mode('overwrite').save('output')
# Filter the DataFrame to only include rows where the name starts with "A"
df = df.filter(df["name"].startswith("A"))

# Add a new column to the DataFrame called "age"
df = df.withColumn("age", df["id"] * 10)

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

```
5
+---+-----+
| id| name|
+---+-----+
|  1|Alice|
```

```
|  2|  Bob|
|  3|Carol|
|  4| Dave|
|  5|  Eve|
+---+-----+

+---+-----+---+
| id| name|age|
+---+-----+---+
|  1|Alice| 10|
+---+-----+---+
```

[ ]: