ex48

August 14, 2022

[ ]:
```
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession

conf = SparkConf().setAppName("ex48")
sc = SparkContext(conf=conf)
ssql = SparkSession.builder.getOrCreate()
```

[2]:
```
inputPath = "data/Ex48/data/"
outputPath = "out48/"
```

[ ]:
```
df = ssql.read.load(
    inputPath,
    format="csv",
    header=True,
    inferSchema=True
)
df.show()
```

[5]:
```
dfNameCountedAgeAvareged = df.groupBy("name").agg({"name":"count", "age":"avg"})
```

[7]:
```
df_filtered = dfNameCountedAgeAvareged.filter("count(name)>=2")
```

[8]:
```
final_df = df_filtered.select("name", "avg(age)")
```

[9]:
```
final_df.write.csv(outputPath, header=False)
```

[ ]:
```
#using SQL
df.createOrReplaceTempView("people")
df_sql = ssql.sql("""
SELECT name, avg(age) as ageavg
FROM people
GROUP BY name
HAVING count(*) >=2
""")
```