babu2015_78

October 3, 2021

[1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import input_file_name

# Assembly or download spark-excel and its dependencies
jars = [
    "/home/quanghgx/notebooks/spark-excel_2.12-3.1.2_0.14.
 ↪0+17-f60f9b12-SNAPSHOT.jar",
    "/home/quanghgx/notebooks/poi-ooxml-schemas-4.1.2.jar",
    "/home/quanghgx/notebooks/commons-collections4-4.4.jar",
    "/home/quanghgx/notebooks/xmlbeans-3.1.0.jar"
]

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.jars", ",".join(jars)) \
    .getOrCreate()
```

[2]:
```
import pyspark

print(pyspark.__file__)
```

```
/home/quanghgx/.local/lib/python3.8/site-packages/pyspark/__init__.py
```

[3]:
```
spark.read.format("excel") \
   .option("header", True) \
   .option("inferSchema", True) \
   .load(f"/home/quanghgx/Downloads/babu2015_78.xlsx") \
   .show(100)
```

```
+--------+----+
|    col1|col2|
+--------+----+
|     cat| 100|
|    lion|  80|
|   whale|  95|
|elephant| 105|
+--------+----+
```

[ ]: