jupyetr与pysaprk联合使用指南

笔记本:	Spark		
创建时间:	2020/1/10 16:02	**更新时间:**	2020/1/10 17:39
作者:	吲哚乙酸		

使用方式1：jupyter中导入pyspark使用

step1: 打开jupyter, 映射已经设置好，端口号10.129.2.155:8899

```
hadoop@slave1:~$ jupyter notebook
```

step2: 新建ipynb文件，输入

```
import os
import sys
spark_name = os.environ.get('SPARK_HOME',None)
if not spark_name:
    raise ValueErrorError('spark环境没有配置好')

sys.path.insert(0,os.path.join(spark_name,'python'))
sys.path.insert(0,os.path.join(spark_name,'python/lib/py4j-0.10.4-src.zip'))
exec(open(os.path.join(spark_name,'python/pyspark/shell.py')).read())
```

如图:

就可以使用了

参考链接
主要是将pyspark中的一些包导入python中

缺点：不能分配核数与内存，都是默认的参数

Name	Cores	Memory per Node
pyspark-shell	144	1024.0 MB

使用方式2：

直接在命令行中启动pyspark，设置启动选项为jupyter

前提anaconda或者jupyter已经安装
在命令行中输入：

```
PYSPARK_DRIVER_PYTHON=~/anaconda3/bin/jupyter-notebook pyspark
```

使用方式3：直接在环境变量中添加（不推荐）

step1:

```
vim ~/.bashrc
```

step2: 添加如下信息
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_PYTHON=/home/hadoop/anaconda3/bin/python3
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
step3:
source ~/.bashrc
step4: 启动
运行pyspark直接启动

```
hadoop@slave1:~$ pyspark
```

或者指定一些：

```
pyspark --master spark://slave:7077 --executor-memory xxxM --total-executor-
cores xx
```

缺点：

因为export了PYSPARK_DRIVER_PYTHON与PYSPARK_DRIVER_PYTHON_OPTS两个环境变量后，非shell的pyspark 生怕认可应用也将使用jupyter-notebook，这必然引起混乱，**所以推荐的还是在pyspark的启动命令中当时指定**
实测，添加到环境变量中后，spark只支持ipynb的文件，因此不用。

参考链接