准备工作：

```
# pyspark
export PYSPARK_DRIVER_PYTHON=jupyter
# export PYSPARK_DRIVER_PYTHON=/usr/bin/python3
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
export PYSPARK_PYTHON=python3
# spark streaming
export PYTHONPATH=$SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.10.7-src.zip:$PYTHONPATH
```

激活环境变量

```
chen@ubuntu:~$ . .bashrc
```

启动hadoop

```
chen@ubuntu:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as chen in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
```

启动spark

```
chen@ubuntu:~$ /apps/spark/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /apps/spark/logs/spark-chen-org.apache.spark.deploy.master.Master-1-ubuntu.out
```

检查

```
chen@ubuntu:~$ jps
12832 DataNode
14977 Worker
12658 NameNode
15028 Jps
14839 Master
13640 NodeManager
13354 ResourceManager
13119 SecondaryNameNode
```

上传数据

```
hadoop fs -mkdir /input/mllib
hadoop fs -put /data/train.tsv /input/mllib/
hadoop fs -put /data/test.tsv /input/mllib/
```

```
chen@ubuntu:/data$ hadoop fs -mkdir /input/mllib
chen@ubuntu:/data$ hadoop fs -put /data/train.tsv /input/mllib/
chen@ubuntu:/data$ hadoop fs -put /data/test.tsv /input/mllib/
```

检查

```
hadoop fs -ls /input/mllib
```

```
chen@ubuntu:/data$ hadoop fs -ls /input/mllib
Found 2 items
-rw-r--r--   1 chen supergroup    9428650 2020-12-15 19:30 /input/mllib/test.tsv
-rw-r--r--   1 chen supergroup   21972916 2020-12-15 19:29 /input/mllib/train.tsv
```

启动pyspark

创建目录~/pyspark-workspace/mllib/，在该目录中启动 Pyspark，创建一个 Notebook。

```
chen@ubuntu:~$ cd -
/home/chen/pyspark-workspace/mllib
chen@ubuntu:~/pyspark-workspace/mllib$ pyspark
[I 19:33:49.214 NotebookApp] Writing notebook server cookie secret to /run/user/1000/jupyter/notebook_cookie_secret
[I 19:33:49.547 NotebookApp] Serving notebooks from local directory: /home/chen/pyspark-workspace/mllib
[I 19:33:49.547 NotebookApp] 0 active kernels
[I 19:33:49.547 NotebookApp] The Jupyter Notebook is running at:
[I 19:33:49.548 NotebookApp] http://10.0.0.135:8888/
[I 19:33:49.548 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
```