Hadoop 实用案例：https://www.cnblogs.com/aolemon/p/13936357.html

https://www.10qianwan.com/articledetail/350061.html

**hadoop streaming 原理**

hadoop 本身是用 java 开发的，程序也需要用 java 编写，但是通过 hadoop streaming，我们可以使用任意语言来编写程序，让 hadoop 运行。

hadoop streaming 就是通过将其他语言编写的 mapper 和 reducer 通过参数传给一个事先写好的 java 程序（hadoop 自带的 *-streaming.jar），这个 java 程序会负责创建 mr 作业，另开一个进程来运行 mapper，将得到的输入通过 stdin 传给它，再将 mapper 处理后输出到 stdout 的数据交给 hadoop，经过 partition 和 sort 之后，再另开进程运行 reducer，同样通过 stdin/stdout 得到最终结果。

**python 的 mapreduce 代码**

因此，使用 python 编写 mapreduce 代码的技巧就在于我们使用了 hadoopstreaming 来帮助我们在 map 和 reduce 间传递数据通过 stdin (标准输入)和 stdout (标准输出).我们仅仅使用 python 的 sys.stdin 来输入数据，使用 sys.stdout 输出数据，这样做是因为 hadoopstreaming 会帮我们办好其他事。

1. word count example

a. 修改文件权限

```
$ chmod +x mapper.py reducer.py
```

b. 上传 input

```
$ hdfs dfs -put *.txt input
```

c. Hadoop-streaming

```
$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar -input input -output output -mapper wc/mapper.py -reducer wc/reducer.py
```

d. get output 到本地

```
$ hdfs dfs -get output output
```

常见问题：

1.当 output 有冲突时：

```
$ hadoop fs -rmr hdfs://localhost:9000/user/jianbo/output
```