在 hadoop 上运行 python&java mapreduce 程序

请去课程中心-资源-小作业下载 MapReduce Assignments(data).zip 解压后查看 wordcount 文件夹
注意：
- 根目录下的 **MapReduce.py wordcount.py** 是独立实现的简易 **MapReduce** 框架以及 **wordcount** 程序 与 **hadoop** 无关。在 **hadoop** 上运行的 **java** 和 **python** 代码分别在 **java python** 目录下
- 在进行以下步骤之前请确保可以顺利执行 **wordcount** 官方样例
- 建议使用 **python** 以 **streaming** 的方式运行。**java** 需要将源码编译打包

挂载本地目录到 docker 容器

docker run 时增加参数 -v :
如： docker run -it -v /home/wangcaimeng/BigData/wordcount:/home/root/a_dir/ suhothayan/hadoop-spark-pig-hive:2.9.2

python

- ```
  hadoop jar /usr/local/hadoop-2.9.2/share/hadoop/tools/lib/hadoop-streamin
  g-2.9.2.jar -D mapred.reduce.tasks=5 -mapper "python <mapper 的 python 脚本
  的绝对路径>" -reducer " python /<reducer 的 python 脚本的绝对路径> " -input <
  输入的路径，注意时 hdfs 文件系统中的路径> -output <输出的路径，注意时 hdfs 文件
  系统中的路径>
  ```
  - 如：

    ```
    hadoop jar /usr/local/hadoop-2.9.2/share/hadoop/tools/lib/hadoop-stre
    aming-2.9.2.jar -D mapred.reduce.tasks=5 -mapper "python /home/root/Bi
    gDataCourse/wordcount/python/wordcountMapper.py " -reducer " python /h
    ```
    ```

```
ome/root/BigDataCourse/wordcount/python/wordcountReducer.py " -input
```

```
/user/root/input/words.json -output /user/root/output/out
```

- getmerge 查看结果

java

- 添加环境变量，因为编译 java 文件时要用的 hadoop 自带的类库,打开 ~ /.bashrc 添加以下内容：

```
HADOOP_HOME=/usr/local/hadoop-2.9.2
```

```
export CLASSPATH=.:$JAVA_HOME/lib/tools.jar:/$JAVA_HOME/lib/dt.jar:$HADOO
```

```
P_HOME/share/hadoop/common/hadoop-common-2.9.2.jar:$HADOOP_HOME/share/had
```

```
oop/mapreduce/hadoop-mapreduce-client-core-2.9.2.jar:$HADOOP_HOME/share/h
```

```
adoop/common/lib/commons-cli-1.2.jar
```

- 将 java 目录下的三个文件编译，并打包成 jar 包（java 基本操作，不会的话就 python 吧）
- hadoop jar xxx.jarargs
- 如：hadoop jar WordCount.jar WordCount.WordCountDirver user/root/input/words.json /user/root/output/out