分布式设计报告

一、问题重述

一 实验目的

1.学习基于MapReduce框架的分布式计算程序设计方法。

2.学习基于Spark框架的分布式计算程序设计方法。

二 实验题目

题目1

输入文件为学生成绩信息，包含了必修课与选修课成绩，格式如下：

班级1, 姓名1, 科目1, 必修, 成绩1 `<br>` （注：`<br>` 为换行符）

班级2, 姓名2, 科目1, 必修, 成绩2 `<br>`

班级1, 姓名1, 科目2, 选修，成绩3 `<br>`

………., ………, ………, ……… `<br>`

编写两个Hadoop平台上的MapReduce程序，分别实现如下功能：

1.计算每个学生必修课的平均成绩。

2.按科目统计每个班的平均成绩。

题目2

输入文件的每一行为具有父子/父女/母子/母女/关系的一对人名，例如：

Tim, Andy `<br>`

Harry, Alice `<br>`

Mark, Louis `<br>`

Andy, Joseph `<br>`

……….., ………… `<br>`

假定不会出现重名现象。

编写Hadoop平台上的MapReduce程序，找出所有具有grandchild-grandparent关系的人名组。

题目3

输入文件为学生成绩信息，包含了必修课与选修课成绩，格式如下：

班级1, 姓名1, 科目1, 必修, 成绩1 `<br>` （注：`<br>` 为换行符）

班级2, 姓名2, 科目1, 必修, 成绩2 `<br>`

班级1, 姓名1, 科目2, 选修，成绩3 `<br>`

………., ………, ………, ……… `<br>`

编写一个Spark程序，同时实现如下功能：

1. 计算每个学生必修课的平均成绩。

2. 统计学生必修课平均成绩在：'90~100,80~89,70~79,60~69和60分以下'这5个分数段的人数。

二、遇到的问题以及解决方案

1、Docker方案的无法实现

最开始我采用了docker去配置环境，但在配置多节点时遇到了没有办法解决的问题：我的ssh无法连接、node节点不断被访问以至于超过负载无法运行，在修改多节点环境后可能损害了单节点的环境，以至于当我再次尝试单节点的环境时仍然报错，无奈之下最终使用了这个在本地的方案。

2、环境配置遇到的问题及解决方案

## 无法解压tar.gz文件？

在打开管理员模式的命令提示符中，使用指令tar -zxvf fileName.tar.gz

## NodeManager启动时遇到错误

NodeManager启动时遇到了一个java.lang.ExceptionInInitializerError

原因是JAVA版本太高了，我用的是java22，请重新下载JAVA8即1.8的版本

GPT解释：特别是Java 9及其以上版本引入了模块系统（Jigsaw项目），导致一些反射操作变得不可访问。而许多Hadoop的库和依赖项（比如Guice和CGLIB）在设计时没有考虑到这些新的限制

## pyspark指令报错

你的Python版本太高了

我这里直接采用了老师的那个版本即python3.6.3这个版本

## 卡在running job后不动了

修改xml里面的配置，包括yarn-site、hdfs-site、core-site、mapred-site，后面会给出我的配置和详细的信息。

## 显示不了中文字符

在控制面板的时钟与区域->区域->管理->更改系统区域设置->选用beta版

重启电脑即可。奇怪的是当我更改重启电脑后就无法打开idea这个软件了，很可能与其中的配置相冲突了

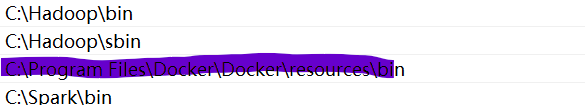
3、环境配置

Hadoop 3.0.0

python 3.6.3

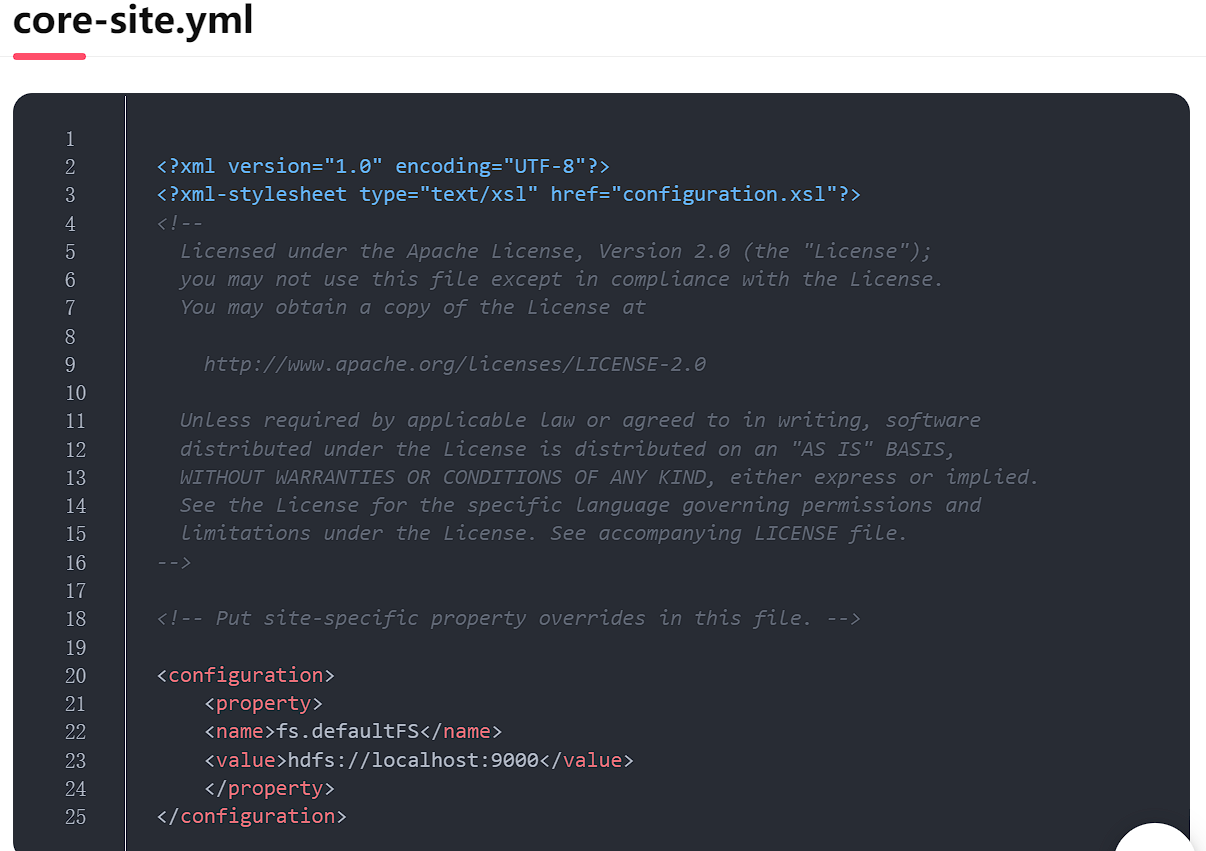
Java 1.8

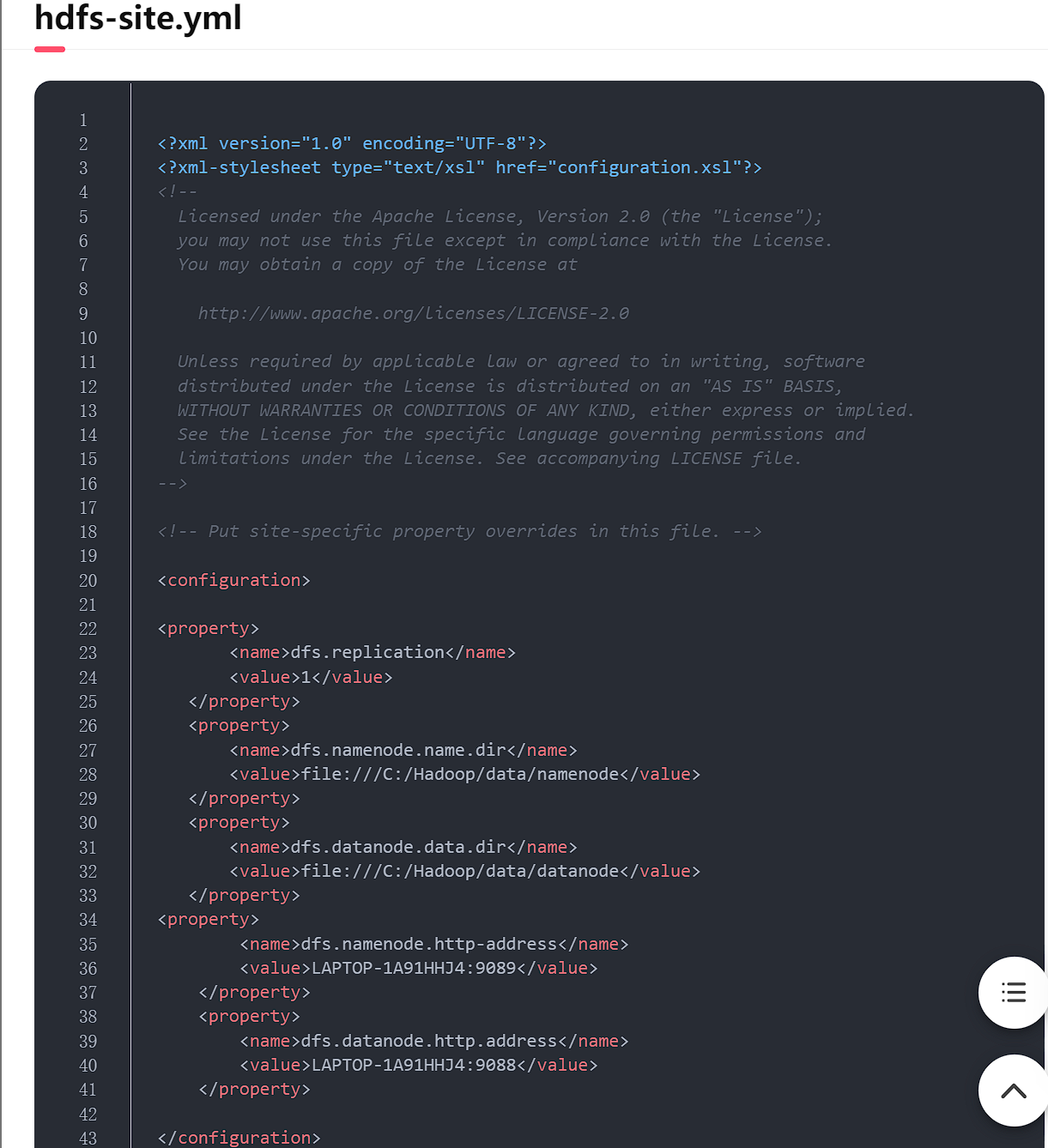
maven 3.9.7

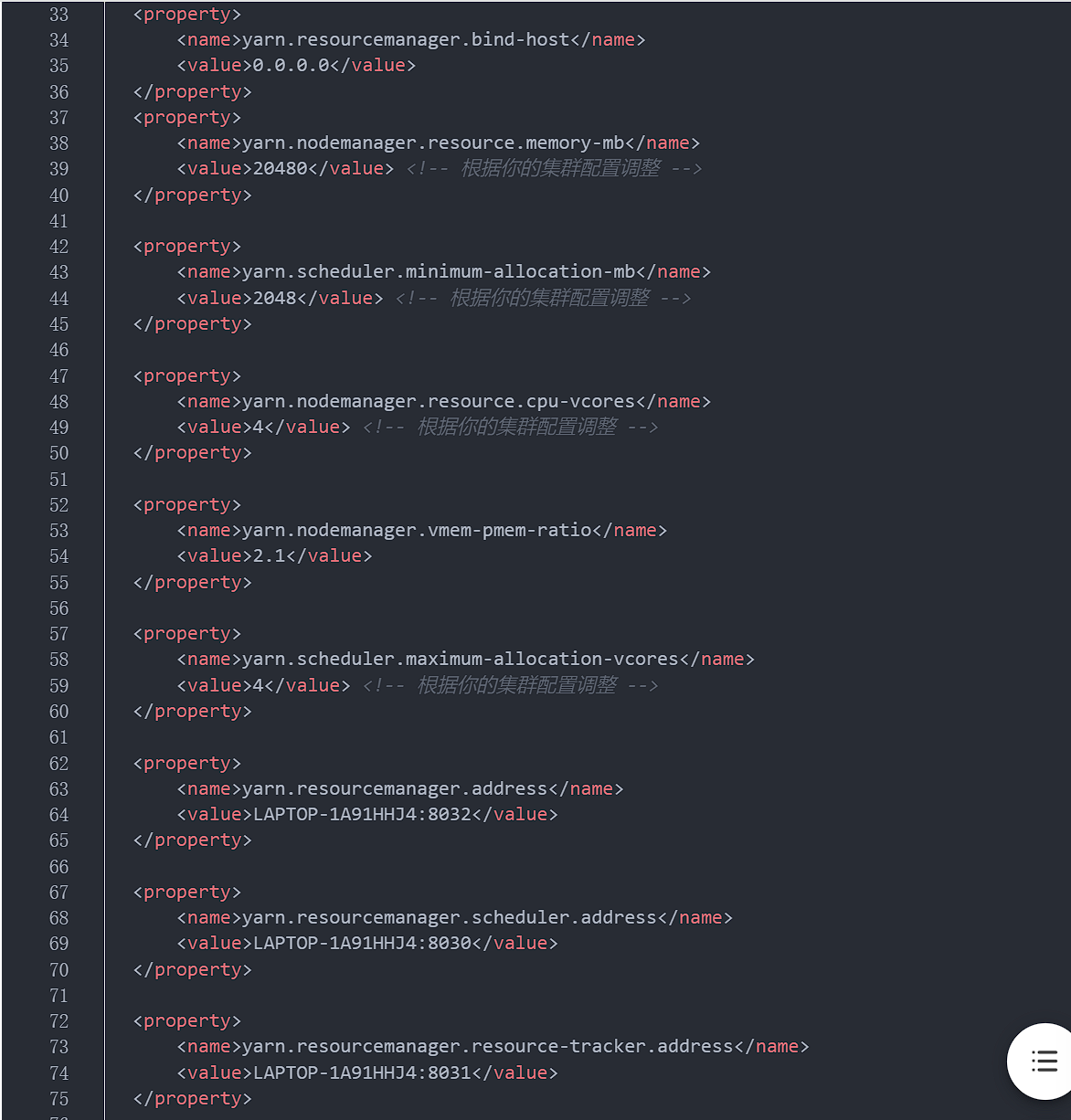
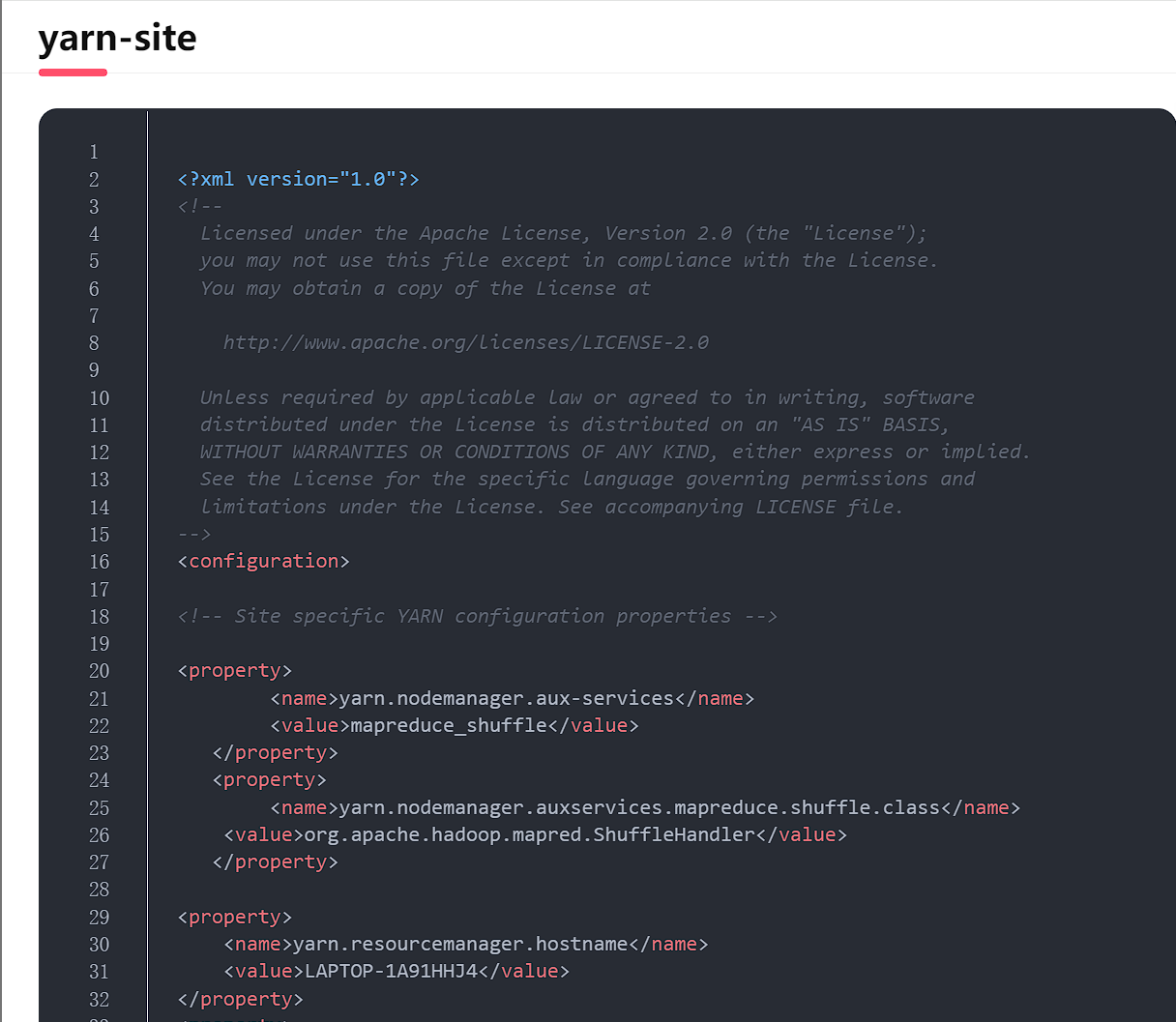
系统变量：

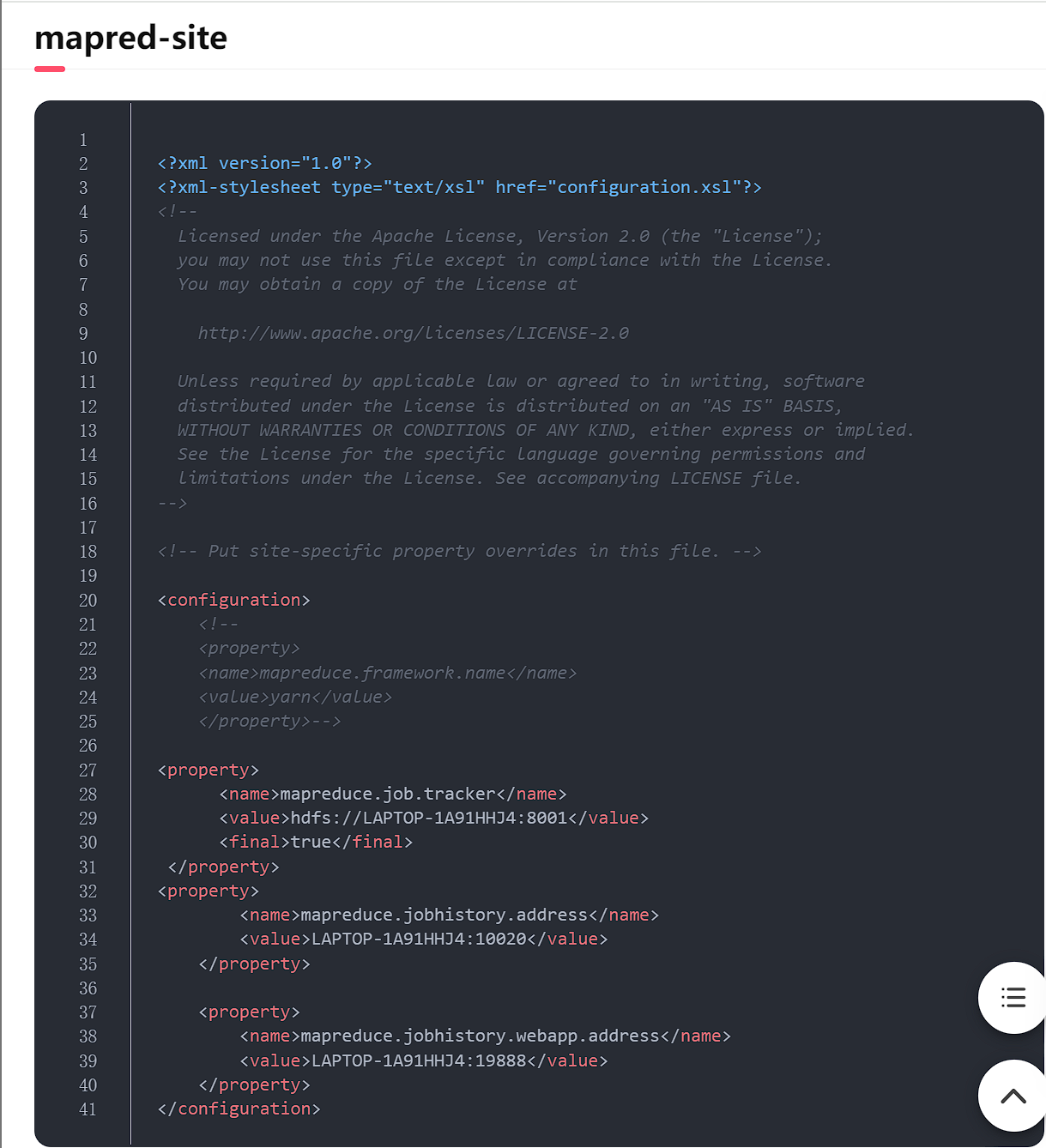
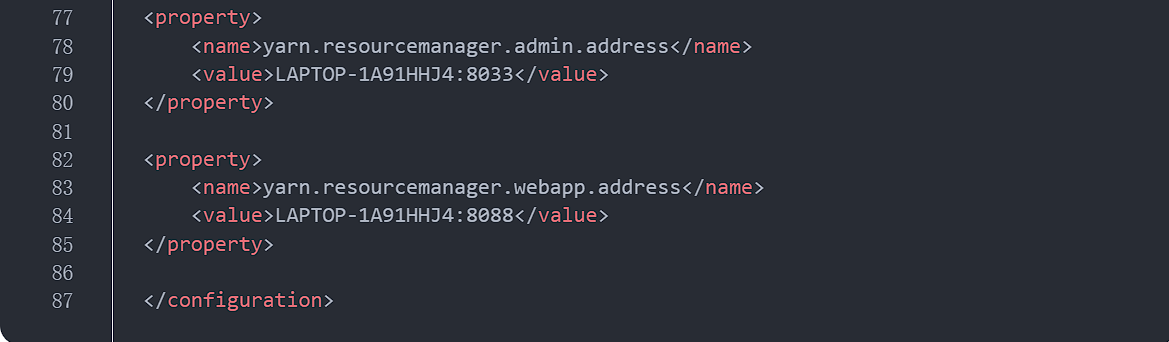
我的主机名是LAPTOP-1A91HHJ4,如果你要查看自己的主机名可在命令指示符中输入'hostname'，即可得到主机名，这是后面更改相关配置的一个需要获取的信息

我这里更改了我的环境配置，即yarn-site、hdfs-site、core-site、mapred-site：









由于我走了很多的弯路，可能这上面的配置多了很多不必要的东西，但一定是可以跑通的。

如果遇到了类似端口占用的报错请直接删除它，例如报错：Port in use:0.0.0.0:8042,首先检查占用端口的进程：  
‘netstat -ano | findstr :8042’  
找到他的进程ID，输入指令:’taskill /PID yourIdNum /F’即可

## Pom环境配置



三、 具体指令

## 打开hadoop

```

start-all

```

## StuAvg

```

mvn clean

mvn package

hadoop fs -mkdir /StuAvg

hadoop fs -put grades.txt /StuAvg

hadoop jar ./target/StuAvg.jar com.ls.StuAvg /StuAvg/grades.txt /output1

hadoop fs -ls /output1

hadoop fs -cat /output1/part-r-00000

hadoop fs -rm -r /output1

```

## ClassAvg

```

mvn clean

mvn package

hadoop fs -mkdir /ClassAvg

hadoop fs -put grades.txt /ClassAvg

hadoop jar ./target/ClassAvg.jar com.ls.ClassAvg /ClassAvg/grades.txt /output2

hadoop fs -ls /output2

hadoop fs -cat /output2/part-r-00000

hadoop fs -rm -r /output2

```

## GrandPC

```

mvn clean

mvn package

hadoop fs -mkdir /GrandPC

hadoop fs -put child-parent.txt /GrandPC

hadoop jar ./target/GrandPC.jar com.ls.GrandPC /GrandPC/child-parent.txt /output3

hadoop fs -ls /output3

hadoop fs -cat /output3/part-r-00000

hadoop fs -rm -r /output3

```

## Spark

```

start-dfs

hadoop fs -mkdir /spark

hadoop fs -put C:\Users\11523\Desktop\Spark\grades.txt /spark

pyspark

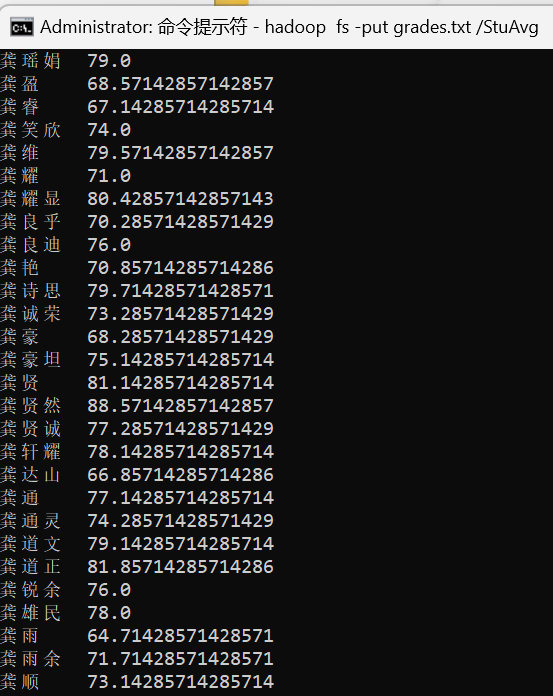
spark-submit Spark.py

hadoop fs -cat /interval\_stu\_nums/part-00000

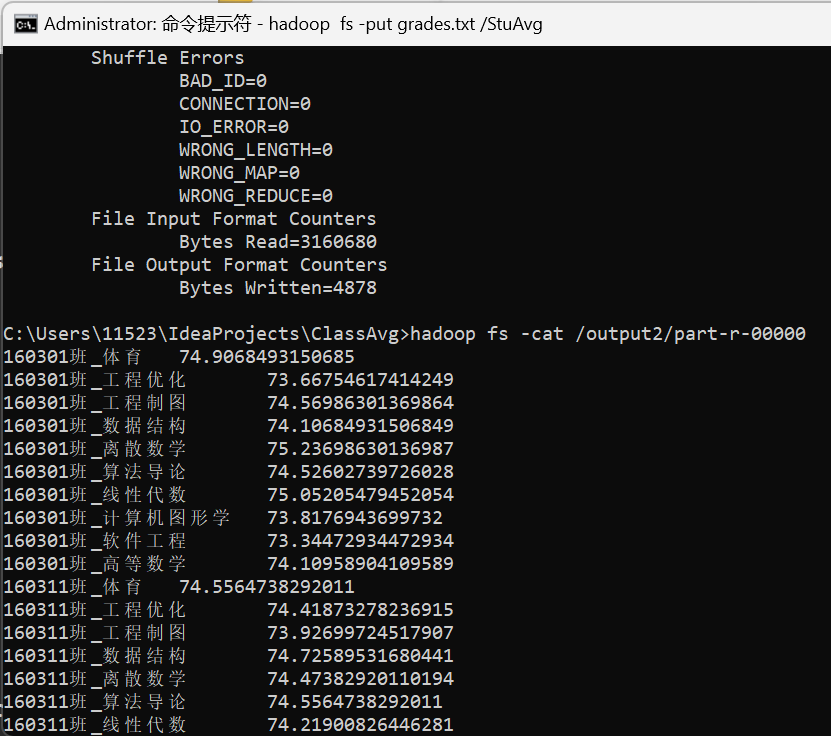
hadoop fs -cat /avg\_grades/part-00000

五、实际效果

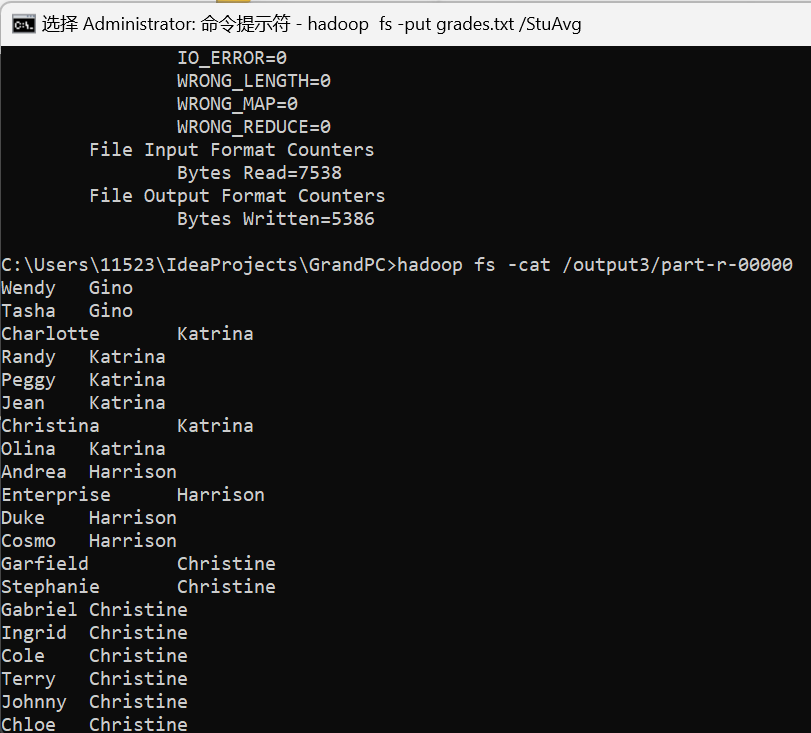
1、StuAvg



2、ClassAvg



3、GrandPC



4、Spark

