第二届“全国大学生大数据技能竞赛”选拔赛：大数据集群搭建+爬取					
序号	名称	分值占比	考核内容	考核技能	建议学习内容
1	基础环境	20%	完成任务书要求的 Linux 基本环境配置。主要考试选手对于 Linux 操作系统的使用及配置方法。	1. 配置 hostname； 2. 安装 net-tools； 3. 配置 firewall； 4. 配置 ntp 服务； 5. 配置 java 环境； 6. 配置 ssh 登陆；	Linux 基础命令(文本操作/目录操作/等)； centos7 防火墙命令； JDK 安装； 环境变量配置； 时间同步/私钥公钥；
2	Hadoop 集群环境搭建	50%	按照任务书要求完成集群搭建，按照正常的大数据集群搭建顺序构建环境，涉及到 Hadoop 的配置情况、配置文件的存在意义以及集群开启方式；	1. 安装 zookeeper； 2. 安装 Hadoop； 3. 安装 Hbase； 4. 安装 MySQL； 5. 安装 Hive；	大数据搭建顺序； 文件的解压安装； vim 操作/文件配置修改； 大数据服务启动顺序； 根据日志查找报错原因；
3	数据爬取	20%	完成相关爬虫环境搭建，按照要求进行对应数据的获取，并保存至指定位置。	1. 编写爬虫代码； 2. 获取数据保存至	Python3.6 的使用； Python 库的使用；

				指定位置； 	熟悉网页结构； 能够独立完成代码编写；
4	Spark 环境搭建	10%	完成任书要求的 Spark 环境安装配置。主要考察选手对于 Spark 组件工具的使用。	1. 安装 Scala 环境； 2. 配置 Spark 组件； 3. 启动 Spark 环境；	软件的安装与配置； 环境变量配置； 根据要求进行 spark 内存等其他配置； 开启服务（了解对应进程)；

其中 maser 主机提供 Python3.6 环境。

支持的库为：

库名	版本号
beautifulsoup4	4.8.0
bs4	0.0.1
certifi	2019.6.16
chardet	3.0.4
html5lib	1.0.1
idna	2.8
lxml	4.4.1
pip	18.1
requests	2.22.0
setuptools	40.6.2
six	1.12.0
soupsieve	1.9.3

urllib3	1.25.3
webencodings	0.5.1

特别提醒：要学会使用日志手段纠错，人为眼检基本找不出错误所在。