**上海市“星光计划”**

**第十届职业院校技能大赛**

“大数据集群搭建与运维”项目

**竞赛试题**

上海市星光计划组委会竞赛办公室

二〇二二年十一月

**竞赛任务及要求**

模块A：操作系统基础环境

Linux是一套免费使用和自由传播的类Unix操作系统，是一个基于POSIX和Unix的多用户、多任务、支持多线程和多CPU的操作系统。伴随着互联网的发展，Linux得到了来自全世界软件爱好者、组织、公司的支持。它除了在服务器操作系统方面保持着强劲的发展势头以外，在个人电脑、嵌入式系统上都有着长足的进步。本项目主要围绕系统基础环境的配置进行考核，内容包括了系统的用户与群组设置、系统权限设置、设置定时任务、系统内压缩文件的使用、find命令、awk命令利用awk获取文件内容、sed命令、grep查找内容过滤、yum配置等知识内容。通过项目实训练习，能够掌握其基本的理论知识和技术原理，加强自身的技能应用能力。

1. 竞赛时间 40 min
2. 竞赛任务及要求

## 1.Linux操作系统中的用户管理、权限管理

1. 创建一个名为admin的组。
2. 创建一个名为apple的用户，以admin作为附加组。
3. 创建banana用户，以admin作为附加组；创建cherry用户，用户不许登录系统，admin不是其附加组。
4. apple，banana，cherry必须以123456作为用户密码。
5. 拷贝文件/root/docker/web.log到/opt/web.log，配置文件/opt/web.log的权限。
6. 文件/opt/web.log的拥有者是root用户
7. 文件/opt/web.log属于admin组
8. 文件/opt/web.log对任何人都不可执行(包括root)，拥有者可读可写，所属组可读可写，其他人仅可读

## 2.文件和目录管理、文件的打包和压缩

1. Zip命令压缩/etc进行存档另存，存储位置为**/opt/etc.zip**
2. 将上一题所压缩的/opt/etc.zip包，指定解压到/opt/test目录中。

## 定时任务定制

1. 每周日凌晨2点30分，运行cp命令(/usr/bin/cp)对/etc/fstab文件进行备份，存储位置为/opt/etc-fstab。
2. 每月1号、10号、22号凌晨4点每10分钟运行cp命令(/usr/bin/cp)对/etc/passwd文件进行备份，存储位置为/opt/etc-passwd。

## 4.文件查找

1. 使用find命令查找/etc目录下所有目录以及子目录下共有多少个passwd文件（按照文件名搜索，不区分文件名大小），统计数量（输出结果wc -l），将命令写入/root/find/01.sh文件中。

模块B：数据库系统配置

本项目主要考察数据库的基本操作，如修改默认密码、赋予权限、创建库、创建表、导入数据、备份数据、清空表数据、恢复表数据等内容，目的使其掌握从事本专业工作所需的操作系统、网络基础、数据库等基础知识。

1. 竞赛时间 40 min
2. 竞赛任务及要求

## 1.MySQL搭建安装

1. 搭建安装mysql，rpm安装。
2. 启动mysql服务。

## 2密码修改

1. 查看默认密码，登陆mysql并修改密码为123456。
2. 创建‘root’@‘%’远程连接用户，密码设置为123456。

## 3.清洗数据，导入csv数据到表中

1. 将/root/docker/web.log日志文件，将第一列IP字段与访问时间字段写入到/root/data.csv文件中（例：172.31.0.1,27/Jul/2022:11:32:37），清洗数据内容缺省。
2. 创建web表，并将整理数据导入到web表中。

## 4.删除数据库

1. 删除demo库。

模块C：Hadoop集群部署

本模块主要是考核学生对Hadoop集群的管理与使用,要求选手能够完成配置基础环境、Hadoop完全分布式集群搭建、集群动态删除DataNode节点、集群调优与运维等操作，综合训练选手的Hadoop集群部署实践技能。

1. 竞赛时间 50 min
2. 竞赛任务及要求

## 1.基础环境配置

1. 主机名与IP映射。
2. SSH免密配置。
3. JAVA搭建配置环境。

## 2.Zookeeper搭建部署

1. 解压Zookeeper压缩包到指定位置。
2. 配置核心文件设置变量。
3. 设置集群列表。
4. 开启Zookeeper服务。

## 3.Hadoop搭建部署

1. 解压Hadoop压缩包到指定位置。
2. 配置核心文件设置变量。
3. 设置节点文件。
4. 最对文件系统进行格式化。
5. 启动Hadoop集群查看各节点服务。

模块D：Hive数据仓库安装

本模块主要是考核学生对Hive数据仓库工具的使用,要求选手能够完成安装数据库、Hive基础环境配置、配置Hive元数据至MySQL、配置Hive客户端、启动Hive等操作，综合训练选手的Hive数据仓库安装实践技能。

1. 竞赛时间 30 min
2. 竞赛任务及要求

## 1.MySQL搭建安装

1. 启动mysql服务。
2. 修改初始密码。
3. 配置远程连接用户和密码。

## 2.Hive搭基础环境配置

1. 解压Hive安装包到指定路径。
2. 配置Hive环境变量。
3. 修改Hive运行环境。
4. 解决jline的版本冲突。

## 3.配置Hive服务端和客户端

1. 驱动JDBC拷贝至hive安装目录。
2. 配置元数据数据存储位置为/user/hive\_remote/warehouse。
3. 配置数据库连接为MySQL。
4. 配置连接JDBC的URL地址主机名及默认端口号3306，数据库为hive，如不存在自行创建，ssl连接方式为false。
5. 配置数据库连接用户。
6. 配置数据库连接密码。

模块E：Spark分析平台搭建

本模块主要是考核学生对Spark分析平台的安装与使用,要求选手能够完成Spark安装包解压、配置环境变量、配置Spark配置文件、修改slaves配置文件、开启集群并验证等操作，综合训练选手的Spark分析平台搭建实践技能。

1. 竞赛时间 30 min
2. 竞赛任务及要求

## 1.Spark集群搭建安装

1. 将Scala安装包解压到指定路径。
2. 配置Scala环境变量。
3. 将Spark安装包解压到指定路径。
4. 配置Spark环境变量。
5. 修改配置文件spark-env.sh，设置JAVA\_HOME、HADOOP\_HOME、HADOOP\_CONF\_DIR、SPARK\_MASTER\_IP、SPARK\_WORKER\_MEMORY。
6. 添加spark从节点文件。

模块F：大数据集群应用与数据分析

本模块主要是考核学生对大数据集群的综合应用与数据分析,要求选手能够使用Hadoop中的MapReduce计算框架、Hive数据仓库中的HQL语言、Spark计算引擎完成日志类数据的离线分析,综合训练选手对Hadoop、Hive、Spark的实际使用技能。

信息流、物流和资金流三大平台是电子商务的三个最为重要的平台。而电子商务信息系统最核心的能力是大数据能力，包括大数据处理、数据分析和数据挖掘能力。无论是电商平台还是在电商平台上销售产品的卖家，都需要掌握大数据分析的能力。越成熟的电商平台，越需要以通过大数据能力驱动电子商务运营的精细化，更好的提升运营效果，提升业绩。

数据集为某购物平台在“双11”之前和之后的过去6个月内的匿名用户的购物日志以及指示它们是否是重复购买者的标签信息。 由于隐私问题，数据采取的方式有偏差，所以这个数据集的统计结果会偏离平台购物的实际情况，但是这不会影响解决方案的适用性。

1. 竞赛时间 50 min
2. 竞赛任务及要求

## 1.电商数据分析项目

1. 构建数据仓库。
2. 开启Hive服务，构建数据表，表的结构按照数据类型进行创建，将数据集完整导入Hadoop系统中，并可使用Hive访问数据。
3. 对本数据集进行分析及解读，通过分析了解数据分布，为后续结果进行展示。
4. 根据用户浏览(点击)最多的品牌，计算该品牌的的收藏购买转化率,结果写入/root/collect\_emption路径下。
5. 查找最活跃用户，求出该用户对应的点击购买转化率最高的品牌信息,并将结果写入/root/click\_emption路径下。