**上海市“星光计划”第十一届职业院校技能大赛**

**（高职组）**

“大数据应用开发”项目

**竞赛样题**

|  |
| --- |
| 上海市星光计划组委会竞赛办公室  2025年三月 |

# **模块一：**操作系统基础环境

## **题目一：**Linux系统基础配置

**题目说明：**

本模块主要是考核学生对Linux操作系统的基础使用,要求选手能够使用对应的命令完成用户管理、权限管理、文件打包与解压缩、计划任务的管理、文件的查找、文件编辑等操作,综合考察选手的操作系统基础环境配置实践技能。

### **任务一：**用户及权限管理

**任务说明：**

1. 创建用户 student，并设置密码。
2. 创建用户组 staff。
3. 将 student 用户添加到 staff 组中。
4. 创建一个文件 /opt/student.txt。
5. 修改文件 /opt/student.txt 的所有者为 student，并将用户组设置为 staff。
6. 修改文件 /opt/student.txt 权限，确保文件对用户可读写执行，对组用户可读执行，对其他用户可读。

**任务要求：**

1. 使用命令创建一个新用户 student，并为其设置密码。
2. 创建一个名为 staff 的用户组。
3. 将 student 用户添加到 staff 用户组中。
4. 创建一个文件 /opt/student.txt，并确保文件存在。
5. 修改文件 /opt/student.txt 的所有者为 student，并将文件所属的用户组更改为 staff。
6. 修改文件 /opt/student.txt 的权限，确保其对用户可读写执行，对组用户可读执行，对其他用户可读。

### **任务二：**文件打包与计划任务管理

**任务说明：**

1. 配置计划任务，使其每天中午12点执行 ls -l /opt/student 命令，并将输出保存到 /opt/daily\_log.txt。
2. 使用 tar 命令将 /opt/student/ 目录压缩为 /opt/student\_home.tar.gz。
3. 解压 /opt/student\_home.tar.gz 文件到 /opt/student\_backup/ 目录。

**任务要求：**

1. 创建 /opt/student/ 目录。
2. 使用 crontab 配置计划任务，使其每天中午12点执行 ls -l /opt/student 命令，并将输出结果保存到 /opt/daily\_log.txt 文件中。
3. 使用 tar 命令将 /opt/student/ 目录打包并压缩为 /opt/student\_home.tar.gz 文件。
4. 解压 /opt/student\_home.tar.gz 文件到 /opt/student\_backup/ 目录。

# **模块二：**大数据集群搭建

## **题目一：**基础配置

**题目说明：**

1. **比赛框架**

本次比赛为分布式集群搭建，共三台节点，其中master作为主节点，slave1、slave2为从节点；

1. **比赛内容**

基础配置：修改主机名、主机映射、时区修改、时间同步、定时任务、免密访问；

JDK安装：环境变量；

Zookeeper部署：环境变量、配置文件zoo.cfg、myid；

Hadoop部署：环境变量、配置文件修改、设置节点文件、格式化、开启集群；

Hive部署：Mysql数据库配置、服务器端配置、客户端配置。

### **任务一：**基础环境配置

**任务说明：**

相关安装包已经存放至环境/hadoop\_packages/目录中

对应ntp和mysql已安装，可直接对其进行操作和配置

1.修改主机名，便于识别节点；

2.工具包已保存在环境中；

3.修改hosts文件，添加集群节点映射，按照给出的节点IP和对应的主机名进行设置；

4.要求各节点时区修改为中国时区（ 中国标准时间CST+8）

5.安装ntp服务，要求主节点master为本地时钟源，从节点设置定时任务同步本地时间；

6.集群中数据传输需要节点之间免密访问，要求设置主节点之间到从节点的免密访问；

7.Hadoop技术基于Java语言，要求本地源下载对应安装包进行安装配置，注意安装路径要求，无需更改文件名，注意添加环境变量。

本环境用于为基础设置部分，用于后续的集群搭建。

**任务要求：**

1. 虚拟机名称修改对应主机名（分别为master、slave1、slave2，使用hostnamectl命令）。
2. 修改host文件添加master、slave1、slave2节点IP与主机名映射（使用内网IP）。
3. 时区更改为上海时间（CST+0800时区）。
4. 环境已经安装NTP，修改master节点NTP配置，设置master为本地时间服务器，屏蔽默认server，服务器层级设为10。
5. 开启NTP服务。
6. 添加定时任务--在早十-晚五时间段内每隔半个小时同步一次本地服务器时间（24小时制、使用用户root任务调度crontab，服务器地址使用主机名）。
7. 主节点生成公钥文件id\_rsa.pub(数字签名RSA，用户root，主机名master)。
8. 建⽴master⾃身使⽤root⽤户ssh访问localhost免密登录。
9. 建⽴master使⽤root⽤户到slave1的ssh免密登录访问。
10. 建⽴master使⽤root⽤户到slave2的ssh免密登录访问。
11. 将jdk安装包解压到/opt/java目录（安装包存放于/hadoop\_packages/，路径自行创建，解压后文件夹为默认名称，其他安装同理）。
12. 配置系统环境变量JAVA\_HOME，同时将JDK安装路径中bin目录加入PATH系统变量，注意生效变量，查看JDK版本。

## **题目二：**Zookeeper搭建

**题目说明：**

Zookeeper是一个分布式服务框架，是Apache Hadoop 的一个子项目，它主要是用来解决分布式应用中经常遇到的一些数据管理问题，如：统一命名服务、状态同步服务、集群管理、分布式应用配置项的管理等。

预装的配置文件zoo\_sample.cfg下面默认有五个属性，分别是：

1. **tickTime**

心跳间隔，单位是毫秒，系统默认是2000毫秒，也就是间隔两秒心跳一次。

tickTime的意义：客户端与服务器或者服务器与服务器之间维持心跳，也就是每个tickTime时间就会发送一次心跳。通过心跳不仅能够用来监听机器的工作状态，还可以通过心跳来控制Flower跟Leader的通信时间，默认情况下FL的会话时长是心跳间隔的两倍。

1. **initLimit**

集群中的follower服务器(F)与leader服务器(L)之间初始连接时能容忍的最多心跳数（tickTime的数量）。

1. **syncLimit**

集群中flower服务器（F）跟leader（L）服务器之间的请求和答应最多能容忍的心跳数。

1. **clientPort**

客户端连接的接口，客户端连接zookeeper服务器的端口，zookeeper会监听这个端口，接收客户端的请求访问，端口默认是2181。

1. **dataDir**

该属性对应的目录是用来存放myid信息跟一些版本，日志，跟服务器唯一的ID信息等。

在集群Zookeeper服务在启动的时候，会回去读取zoo.cfg这个文件，从这个文件中找到这个属性然后获取它的值也就是dataDir 的路径，它会从这个路径下面读取myid这个文件，从这个文件中获取要启动的当前服务器的地址。

### **任务一：**Zookeeper搭建

**任务要求：**

1. 将zookeeper安装包解压到指定路径/opt/zookeeper（安装包存放于/hadoop\_packages/）。
2. 配置系统变量ZOOKEEPER\_HOME，同时将Zookeeper安装路径中bin目录加入PATH系统变量，注意生效变量。
3. Zookeeper的默认配置文件为Zookeeper安装路径下conf/zoo\_sample.cfg，将其修改为zoo.cfg。
4. 设置数据存储路径(dataDir)为/opt/zookeeper/zkdata。
5. 设置日志文件路径(dataLogDir)为/opt/zookeeper/zkdatalog。
6. 设置集群列表（要求master为1号服务器，slave1为2号服务器，slave2为3号服务器）。
7. 创建所需数据存储文件夹、日志存储文件夹。
8. 数据存储路径下创建myid，写入对应的标识主机服务器序号。
9. 启动服务，查看进程QuorumPeerMain是否存在。
10. 查看各节点服务器角色是否正常(leader/follower)。

## **题目三：**Hadoop集群搭建

**题目说明：**

Hadoop是由Java语言编写的，在分布式服务器集群上存储海量数据并运行分布式分析应用的开源框架，其核心部件是HDFS与MapReduce。

HDFS是一个分布式文件系统：引入存放文件元数据信息的服务器Namenode和实际存放数据的服务器Datanode，对数据进行分布式储存和读取。

MapReduce是一个计算框架：MapReduce的核心思想是把计算任务分配给集群内的服务器里执行。通过对计算任务的拆分（Map计算/Reduce计算）再根据任务调度器（JobTracker）对任务进行分布式计算。

### **任务一：**Hadoop完全分布式集群搭建

**任务要求：**

1. 将Hadoop安装包解压到指定路径/opt/hadoop（安装包存放于/hadoop\_packages/）。
2. 配置环境变量HADOOP\_HOME，将Hadoop安装路径中bin目录和sbin目录加入PATH系统变量，注意生效变量。
3. 配置Hadoop运行环境JAVA\_HOME。
4. 设置全局参数，指定HDFS上NameNode地址为master,端口默认为9000。
5. 指定临时存储目录为本地/opt/hadoop/tmp(要求为绝对路径，下同)。
6. 设置HDFS参数，指定备份文本数量为2。
7. 设置HDFS参数，指定NN存放元数据信息路径为本地/opt/hadoop/tmp/namenode；指定DN存放元数据信息路径为本地/opt/hadoop/tmp/datanode(要求为绝对路径)。
8. 设置HDFS参数，关闭hadoop集群权限校验（安全配置），允许其他用户连接集群；指定datanode之间通过域名方式进行通信。
9. 设置YARN运行环境$JAVA\_HOME参数。
10. 设置YARN核心参数，指定ResourceManager进程所在主机为master，端口为18141;指定mapreduce 获取数据的方式为mapreduce\_shuffle。
11. 设置计算框架参数，指定MR运行在yarn上。
12. 设置节点文件，要求master为主节点； slave1、slave2为子节点。
13. 对文件系统进行格式化。
14. 启动Hadoop集群查看各节点服务。
15. 查看集群运行状态是否正常。

## **题目四：**Hive集群搭建

**题目说明:**

1. **比赛框架**

本次比赛为分布式集群搭建，共三台节点，其中master作为主节点，slave1、slave2为从节点；

1. **比赛内容**

基础配置：修改主机名、主机映射、时区修改、时间同步、定时任务、免密访问；

JDK安装：环境变量；

Zookeeper部署：环境变量、配置文件zoo.cfg、myid；

Hadoop部署：环境变量、配置文件修改、设置节点文件、格式化、开启集群；

Hive部署：Mysql数据库配置、服务器端配置、客户端配置安装数据库

### **任务一：**安装数据库

**任务说明：**

相关安装包已经存放至环境/hadoop\_packages/中，对应ntp和mysql已安装，可直接对其进行操作和配置。

**任务要求：**

1. slave2环境中已经安装mysql-community-server，关闭mysql开机自启服务。
2. 开启MySQL服务。
3. 判断mysqld.log日志下是否生成初临时密码。
4. 设置mysql数据库本地root用户密码为123456。

### **任务二：**Hive基础环境配置

**任务说明：**

Hive是基于Hadoop的一个数据仓库工具，用来进行数据提取、转化、加载，这是一种可以存储、查询和分析存储在Hadoop中的大规模数据的机制。Hive数据仓库工具能将结构化的数据文件映射为一张数据库表，并提供SQL查询功能，能将SQL语句转变成MapReduce任务来执行。

1. 将指定版本的Hive安装包解压到指定路径，添加系统并生效；
2. 修改Hive运行环境；
3. 由于客户端需要和Hadoop通信，为避免jline版本冲突问题，将Hive中lib/jline-2.12.jar拷贝到Hadoop中，保留高版本。

**任务要求：**

1. 将Hive安装包解压到指定路径/opt/hive（安装包存放于/hadoop\_packages/）。
2. 配置环境变量HIVE\_HOME,将Hive安装路径中的bin目录加入PATH系统变量，注意生效变量。
3. 修改HIVE运行环境，配置Hadoop安装路径HADOOP\_HOME。
4. 修改HIVE运行环境，配置Hive配置文件存放路径HIVE\_CONF\_DIR。
5. 修改HIVE运行环境，配置Hive运行资源库路径HIVE\_AUX\_JARS\_PATH。
6. 解决jline的版本冲突，将$HIVE\_HOME/lib/jline-2.12.jar同步至$HADOOP\_HOME/share/hadoop/yarn/lib/下。

### **任务三：**配置HIVE元数据至MySQL

**任务说明：**

slave1作为服务器端需要和Mysql通信，所以服务端需要将Mysql的依赖包放在Hive的lib目录下。mysql-connector-java是MySQL的JDBC驱动包，用JDBC连接MySQL数据库时必须使用该jar包。

**任务要求：**

1. 驱动JDBC拷贝至hive安装目录对应lib下（依赖包存放于/hadoop\_packages/）。
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。

### **任务四：**配置HIVE客户端

**任务说明：**

1. master作为客户端,可进入终端进行操作。
2. 关闭本地模式。
3. 将hive.metastore.uris指向metastore服务器URL。

**任务要求：**

1. 配置元数据存储位置为/user/hive\_remote/warehouse。
2. 关闭本地metastore模式。
3. 配置指向metastore服务的主机为slave1，端口为9083。

### **任务五：**启动Hive

**任务说明：**

1. 服务器端初始化数据库，并启动metastore服务。
2. 客户端开启Hive client，即可根据要求创建相关数据操作。

**任务要求：**

1. 服务器端初始化数据库，启动metastore服务。
2. 客户端开启进入hive，创建hive数据库。

## **题目五：**Spark搭建

**题目说明：**

Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，换句话说，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。

Spark 是在 Scala 语言中实现的，它将 Scala 用作其应用程序框架。与 Hadoop 不同，Spark 和 Scala 能够紧密集成，其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集。

### **任务一：**Spark集群搭建

**任务说明：**

Spark是Hadoop的子项目。在相关环境中将Spark安装到基于Linux的系统中。

**任务要求：**

1. 将Spark安装包解压到指定路径。/opt/spark/（安装包存放于/hadoop\_packages/）。
2. 文件/etc/profile中配置环境变量SPARK\_HOME，将Spark安装路径中的bin目录加入PATH系统变量，注意生效变量。
3. 修改配置文件spark-env.sh，设置主机节点为master。
4. 修改配置文件spark-env.sh，设置java安装路径。
5. 修改配置文件spark-env.sh，设置节点内存为8g。
6. 修改配置文件spark-env.sh，设置hadoop安装目录、hadoop集群的配置文件的目录。
7. 修改slaves文件，添加spark从节点slave1、slave2。
8. 开启集群，查看各节点进程(主节点进程为Master，子节点进程为Worker)。

# **模块三：**数据预处理与存储

## **题目一：**缺失值处理-数据填充

**题目说明：**

根据给出的数据集进行相关字段填充操作，限定填充方式，填充有缺失值的字段，将填充后数据存放在指定位置，可能是HDFS或导入至Hive。

### **任务一：**数据填充

**任务要求：**

1. 编写代码读取给定数据集`titanic3.csv`。
2. 给定数据集的部分字段有缺失值，找出对应字段。
3. 结合缺失值字段数据的特点选择更优的数据填充方式进行填充。
4. 将完成填充后的数据导出至HDFS路径/3\_imputer。

## **题目二：**数据清洗

**题目说明：**

根据给出的数据集进行相关字段清洗操作，限定清洗方式，提取所需字段，将提取后的数据存放在指定位置，可能是HDFS或导入至Hive。。

### **任务一：**数据清洗

**任务要求：**

1. 编写代码清洗给定日志数据文件。
2. 原始数据为日志文件，需要按清洗规则提取其中的IP地址字段与访问时间字段。
3. 将清洗后的数据文件导出至hive，新建hive表log并导入数据。

# **模块四：**数据分析

## **题目一：**Hadoop数据分析

### **任务一：**MapReduce实现词频统计

**任务说明：**

单词计数是最简单也是最能体现MapReduce思想的程序之一，可以称为MapReduce版"Hello World"，该程序的完整代码可以在Hadoop安装包的"src/examples"目录下找到。单词计数主要完成功能是：统计一系列文本文件中每个单词出现的次数。

**任务要求：**

1. 将数据word\_count\_data.txt上传至HDFS文件系统/input/下，注意自行创建目录。
2. 编写程序或使用Hadoop自带的开源jar包对数据进行wordcount操作，结果保存至hdfs:/output/part--r-00000。

## **题目二：**Hive数据分析

### **任务一：**车辆GPS位置信息分析

**任务说明：**

现有数据为江苏省南京市出租车从2010年9月1日到2日的GPS位置数据，包含了(7,726)辆出租车的共(33,042,225)条位置记录，可以适用于大数据分析和挖掘。可通过统计出租车的分布密度、速度和方向，分析交通高峰时段的路网压力。根据出租车速度和经纬度信息，可以识别哪些路段容易出现拥堵。统计载客和空车的起始位置，可识别热门上下客点和需求量较大的区域。根据出租车高频上下客点和时段，识别公共交通覆盖不足的区域，给出运营思路和方法上的建议，规划补充措施。

**任务要求：**

1. 将本地数据文件gps.csv导入到Hive上，表名指定为gps。
2. 统计每小时出租车的载客情况：统计每小时总共的独立出租车数量以及在该小时载过客的独立出租车数量，并计算出载过客的比例。结果写入本地/opt/4-2-1.csv。
3. 统计每小时载客出租车的平均车速。结果写入本地/opt/4-2-2.csv。
4. 统计每小时载客出租车的利用率：统计每小时所有出租车载客的总时间和行驶的总时间，并计算出载客时间的比例，即为利用率。结果写入本地/opt/4-2-3.csv。
5. 统计各出租车分别在3时、8时和13时第一次同步数据时的载客情况。结果分别写入本地/opt/4-2-4-1.csv、/opt/4-2-4-2.csv、/opt/4-2-4-3.csv。
6. 统计各载客出租车在3时、8时和13时第一次同步数据时的瞬时速度。结果分别写入本地/opt/4-2-5-1.csv、/opt/4-2-5-2.csv、/opt/4-2-5-3.csv。

## **题目三：**机器学习建模

**题目说明：**

1. 使用机器学习算法完成基本建模任务

基本任务需要我们处理数据，观察特征，完成建模并进行可视化展示分析。

1. 观察数据量与特征个数对结果影响

在保证算法一致的前提下，加大数据个数，观察结果变换。重新考虑特征工程，引入新特征后观察结果走势。

1. 对随机森林算法进行调参，找到最合适的参数

掌握机器学习中经典调参方法，对当前模型进行调节。

### **任务一：**电话营销响应预测建模

**任务说明：**

1. 使用决策树和梯度提升机算法完成基本建模任务

基本任务需要我们处理数据，观察特征，完成建模并进行可视化展示分析。

1. 观察数据量与特征个数对结果影响

在保证算法一致的前提下，加大数据个数，观察结果变换。重新考虑特征工程，引入新特征后观察结果走势。

在数据分析和特征提取的过程中，我们的出发点都是尽可能多的选择有价值的特征，因为起始阶段我们能得到的信息越多，之后建模可以利用的信息也是越多的。

**任务要求:**

1. 加载数据，并查看前5行数据。
2. 探索性分析：查看各变量的分布和缺失值情况。
3. 划分训练和测试集。
4. 类型变量处理。
5. 特征选择。
6. 建立决策树模型。
7. 超参数调优。
8. 模型训练。
9. 在测试集上做模型测试，评估模型性能。
10. 绘制ROC曲线。
11. 建立梯度提升机模型并重复调优和评估步骤。

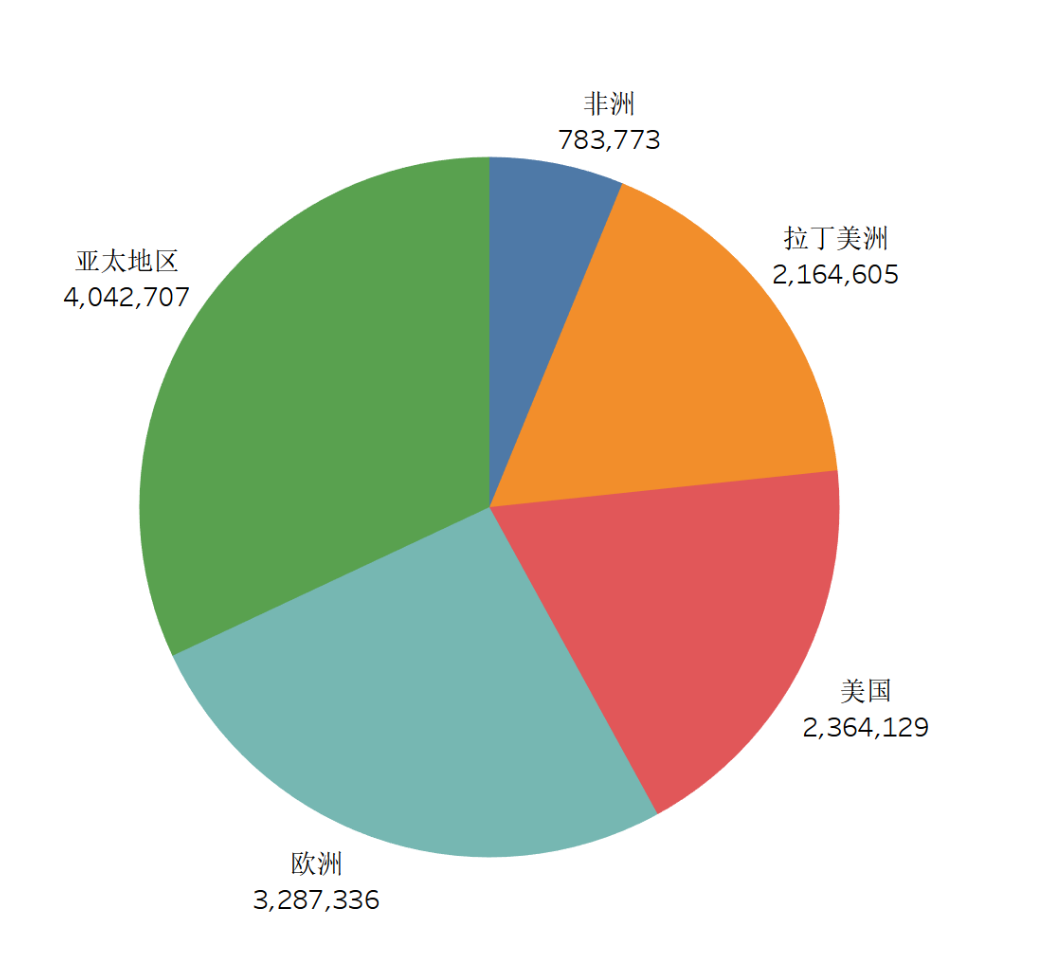
# **模块五：**数据可视化

## **题目一：**数据可视化

### **任务一：**全球大型超市订单数据可视化

**任务说明：**

导入全球大型超市订单数据，实现如下图表，分析该全球超市在各市场的销售额占比情况。



**任务要求：**

1. 读取全球大型超市订单数据。
2. 数据预处理：将原始数据选择合适字段处理成输入形式。
3. 绘制饼图。
4. 设置图标题与图例。
5. 分析各市场占比情况。