**2023年全国职业院校技能大赛**

**赛题第01套**

**赛项名称： 大数据应用开发**

**英文名称：** Big Data Application Development

**赛项组别： 高等职业教育组**

**赛项编号： GZ033**

**背景描述**

大数据时代背景下，电商经营模式发生很大改变。在传统运营模式中，缺乏数据积累，人们在做出一些决策行为过程中，更多是凭借个人经验和直觉，发展路径比较自我封闭。而大数据时代，为人们提供一种全新的思路，通过大量的数据分析得出的结果将更加现实和准确。商家可以对客户的消费行为信息数据进行收集和整理，比如消费者购买产品的花费、选择产品的渠道、偏好产品的类型、产品回购周期、购买产品的目的、消费者家庭背景、工作和生活环境、个人消费观和价值观等。通过数据追踪，知道顾客从哪儿来，是看了某网站投放的广告还是通过朋友推荐链接，是新访客还是老用户，喜欢浏览什么产品，购物车有无商品，是否清空，还有每一笔交易记录，精准锁定一定年龄、收入、对产品有兴趣的顾客，对顾客进行分组、标签化，通过不同标签组合运用，获得不同目标群体，以此开展精准推送。

因数据驱动的零售新时代已经到来，没有大数据，我们无法为消费者提供这些体验，为完成电商的大数据分析工作，你所在的小组将应用大数据技术，以Scala作为整个项目的基础开发语言，基于大数据平台综合利用Hive、Spark、Flink、Vue.js等技术，对数据进行处理、分析及可视化呈现，你们作为该小组的技术人员，请按照下面任务完成本次工作。

## 任务A：大数据平台搭建（容器环境）（15分）

**环境说明：**

|  |
| --- |
| **服务端登录地址详见各任务服务端说明。**  **补充说明：**宿主机及各容器节点可通过Asbru工具或SSH客户端进行SSH访问。 |

### 子任务一：Hadoop 完全分布式安装配置

本任务需要使用root用户完成相关配置，安装Hadoop需要配置前置环境。命令中要求使用绝对路径，具体要求如下:

1. 从宿主机/opt目录下将文件hadoop-3.1.3.tar.gz、jdk-8u212-linux-x64.tar.gz复制到容器Master中的/opt/software路径中（若路径不存在，则需新建），将Master节点JDK安装包解压到/opt/module路径中(若路径不存在，则需新建)，将JDK解压命令复制并粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
| tar -zxvf jdk-8u212-linux-x64.tar.gz -C /opt/module/ |

1. 修改容器中/etc/profile文件，设置JDK环境变量并使其生效，配置完毕后在Master节点分别执行“java -version”和“javac”命令，将命令行执行结果分别截图并粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
| Java -version    Javac |

1. 请完成host相关配置，将三个节点分别命名为master、slave1、slave2，并做免密登录，用scp命令并使用绝对路径从Master复制JDK解压后的安装文件到slave1、slave2节点（若路径不存在，则需新建），并配置slave1、slave2相关环境变量，将全部scp复制JDK的命令复制并粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
| Slave1节点：    Slave2 节点： |

1. 在Master将Hadoop解压到/opt/module(若路径不存在，则需新建)目录下，并将解压包分发至slave1、slave2中，其中master、slave1、slave2节点均作为datanode，配置好相关环境，初始化Hadoop环境namenode，将初始化命令及初始化结果截图（截取初始化结果日志最后20行即可）粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
| 初始化命令：  Hdfs namenode -format  初始化截图： |

1. 启动Hadoop集群（包括hdfs和yarn），使用jps命令查看Master节点与slave1节点的Java进程，将jps命令与结果截图粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下。

|  |
| --- |
|  |

### 子任务二：Spark on Yarn安装配置

本任务需要使用root用户完成相关配置，已安装Hadoop及需要配置前置环境，具体要求如下：

1. 从宿主机/opt目录下将文件spark-3.1.1-bin-hadoop3.2.tgz复制到容器Master中的/opt/software（若路径不存在，则需新建）中，将Spark包解压到/opt/module路径中(若路径不存在，则需新建)，将完整解压命令复制粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
| tar -zxvf spark-3.1.1-bin-hadoop3.2.tgz -C /opt/module/ |

1. 修改容器中/etc/profile文件，设置Spark环境变量并使环境变量生效，在/opt目录下运行命令spark-submit --version，将命令与结果截图粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
|  |

1. 完成on yarn相关配置，使用spark on yarn 的模式提交$SPARK\_HOME/examples/jars/spark-examples\_2.12-3.1.1.jar 运行的主类为org.apache.spark.examples.SparkPi，将运行结果截图粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下（截取Pi结果的前后各5行）。

（运行命令为：spark-submit --master yarn --class org.apache.spark.examples.SparkPi $SPARK\_HOME/examples/jars/spark-examples\_2.12-3.1.1.jar）

|  |
| --- |
|  |

### 子任务三：HBase分布式安装配置

本任务需要使用root用户完成相关配置，安装HBase需要配置Hadoop和ZooKeeper等前置环境。命令中要求使用绝对路径，具体要求如下:

1. 从宿主机/opt目录下将文件apache-zookeeper-3.5.7-bin.tar.gz、hbase-2.2.3-bin.tar.gz复制到容器Master中的/opt/software路径中（若路径不存在，则需新建），将ZooKeeper、HBase安装包解压到/opt/module目录下，将HBase的解压命令复制并粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；

|  |
| --- |
| Hbase 解压命令：  tar -zxvf hbase-2.2.3-bin.tar.gz -C /opt/module/ |

1. 完成ZooKeeper相关部署，用scp命令并使用绝对路径从容器master复制HBase解压后的包分发至slave1、slave2中，并修改相关配置，配置好环境变量，在容器Master节点中运行命令hbase version，将全部复制命令复制并将hbase version命令的结果截图粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下;

|  |
| --- |
| HBase解压后的包分发至slave1  scp -r /opt/module/apache-zookeeper-3.5.7-bin slave1:/opt/module/  HBase解压后的包分发至slave2  scp -r /opt/module/apache-zookeeper-3.5.7-bin slave2:/opt/module/  hbase version |

1. 启动HBase后在三个节点分别使用jps命令查看，并将结果分别截图粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下；正常启动后在hbase shell中查看命名空间，将查看命名空间的结果截图粘贴至客户端桌面【Release\任务A提交结果.docx】中对应的任务序号下。

|  |
| --- |
| Jps master    Jps slave1    Jps slave2    命名空间 |