# 附件2：

# 大数据技术与应用赛项竞赛试题高职组（样卷）

# 竞赛时间、内容及总成绩

（一）竞赛时间

竞赛时间共为6小时，参赛队自行安排任务进度，休息、饮水、如厕等不设专门用时，统一含在竞赛时间内。

（二）竞赛内容概述

|  |  |  |
| --- | --- | --- |
| **序号** | **任务名称** | **具体内容** |
| 任务一 | 大数据平台环境搭建 | 按照大数据分析平台需求，需要完成Hadoop完全分布式、Spark安装配置、Flink安装配置、Hive安装配置、Kafka安装配置、Flume安装配置 |
| 任务二 | 离线数据抽取 | 按照要求使用Scala语言完成特定函数的编写，使用Spark抽取MySQL指定数据表中的新增的数据到ODS层的指定的分区表中 |
| 任务三 | 离线数据统计 | 使用Scala语言编写程序获取ODS层指定分区表中的数据进行清洗，并完成销量前5的商品统计、某月的总销售额统计、指定月份的销售额统计、各用户在线总时长统计，并将统计后的数据存入MySQL数据库中 |
| 任务四 | 数据采集与实时计算 | 启动业务系统，按照要求使用Flume将用户操作日志采集并存入Kafka中并使用Flink、Scala消费Kafka中的数据将其进行聚合计算出商城在线人数，将结果存入Redis中，并统计该系统的UV与PV将结果存入MySQL中 |
| 任务五 | 数据可视化 | 编写前端Web界面，调用后台数据接口，使用Vue.js、Echarts完成数据可视化 |
| 任务六 | 综合分析报告 | 根据要求编写综合分析报告 |

（三）竞赛总成绩

“大数据技术与应用”赛项竞赛总成绩为100分，其中包含赛场职业素养5分。

# 任务须知

1. 每组参赛队分配一台竞赛服务器、三台客户机，拥有独立IP组。
2. 本次比赛采用统一网络环境比赛，请不要随意更改客户端的网络地址信息，对于更改客户端信息造成的问题，由参赛选手自行承担比赛损失；
3. 请不要恶意破坏竞赛环境，对于恶意破坏竞赛环境的参赛者，组委会根据其行为予以处罚直至取消比赛资格。
4. 比赛过程中及时保存相关文档。
5. 比赛相关文档中不能出现参赛学校名称和参赛选手名称，以赛位号（工位号）代替。
6. 参赛选手请勿删除模板内容，若因删除导致任何问题后果自负。
7. 若同一文档由不同选手完成，须将文档合并后作为最终结果提交到U盘中。
8. 比赛中出现各种问题及时向现场裁判举手示意，不要影响其他参赛队比赛。

# 任务说明

本项目要求完成离线电商数据统计分析，完成大数据平台环境搭建、离线数据抽取、离线数据统计、数据采集与实时计算、数据可视化及综合分析报告编写等工作。

提供的相关资源包括：

1. 大数据环境搭建中需要用到的组件安装包
2. 电商相关数据
3. 大数据分析集群环境
4. 离线数据抽取开发环境
5. 离线数据统计开发环境
6. 数据采集与实时计算开发环境
7. 数据可视化开发环境
8. 综合分析报告文档模板

任务一：大数据平台环境搭建

按照大数据分析平台需求，需要完成Hadoop完全分布式、Spark安装配置、Flink安装配置、Hive安装配置、Kafka安装配置、Flume安装配置。

任务二：离线数据抽取

按照要求使用Scala语言完成特定函数的编写，使用Spark抽取MySQL指定数据表中的新增的数据到ODS层的指定的分区表中。

任务三：离线数据统计

使用Scala语言编写程序获取ODS层指定分区表中的数据进行清洗，并完成销量前5的商品统计、某月的总销售额统计、指定月份的销售额统计、各用户在线总时长统计，并将统计后的数据存入MySQL数据库中。

任务四：数据采集与实时计算

启动业务系统，按照要求使用Flume将用户操作日志采集并存入Kafka中并使用Flink、Scala消费Kafka中的数据将其进行聚合计算出商城在线人数，将结果存入Redis中，并统计该系统的UV与PV将结果存入MySQL中。

任务五：数据可视化

编写前端Web界面，调用后台数据接口，使用Vue.js、Echarts完成数据可视化。

任务六：综合分析报告

根据项目要求，完成综合分析报告编写。

# 四、竞赛结果提交要求

（一）提交方式

任务成果需拷贝至提供的U盘中。在U盘中以XX工位号建一个文件夹（例如01），将所有任务成果文档保存至该文件夹中。

（二）文档要求

竞赛提交的所有文档中不能出现参赛队信息和参赛选手信息，竞赛文档需要填写参赛队信息时以工位号代替（XX代表工位号）。