**2018年全国职业院校技能大赛**

**赛项申报书**

赛项名称：大数据处理与分析

赛项类别：常规赛项☑ 行业特色赛项□

赛项组别：中职组□ 高职组☑

涉及的专业大类/类：电子信息大类

方案设计专家组组长：

手机号码：

方案申报单位（盖章）：全国高等院校计算机基础教育研究会

方案申报负责人：

方案申报单位联络人：

联络人手机号码：

电子邮箱：

通讯地址：

邮政编码：

申报日期：2017.9.1

**2018年全国职业院校技能大赛**

**竞赛项目方案**

## **一、赛项名称**

（一）赛项名称

大数据处理与分析

1. 压题彩照



（三）赛项归属产业类型

信息技术产业——战略新兴产业

（四）赛项归属专业大类

1、61电子信息大类

6102计算机类

本赛项主要覆盖专业如下：

**610215 大数据技术与应用**

610201 计算机应用技术

610205 软件技术

610206 软件与信息服务

610203 计算机信息管理

610213 云计算技术与应用

610202 计算机网络技术

610204 计算机系统与维护

610208 嵌入式技术与应用

610212 移动应用开发

2、63财经商贸大类

6308电子商务类

本赛项主要覆盖专业如下：

**630804 商务数据分析与应用**

## **赛项申报专家组**

## **赛项目的**

2016年，习近平总书记指出，“大数据战略作为‘十三五’期间的十四大国家战略之一，是中国经济发展新的驱动力。” 至此，大数据战略上升到国家战略。

大数据，已经成为当前社会的热词。受到我国政府的高度重视和大力支持。2015年8月31日，国务院印发《促进大数据发展行动纲要》（国发〔2015〕50号），2016年3月国家将大数据战略纳入十三五规划。同时各地出台政策 支持大数据产业发展。

2016年9月,教育部研究决定正式批准高职大数据技术与应用专业（专业代码：610215）（《普通高等学校高等职业教育（专科）专业目录2016年增补专业》）；2017年教育部也正式批准了本科专业数据科学与大数据技术（专业代码：080910T）。到2017年申请大数据专业的院校成井喷式的增长，据不完全统计目前已有近100所高职院校均已在开设大数据专业或在计算机专业开设大数据方向，像江苏、湖南、福建等省的高职院校已与企业联合开设大数据学院，共同培养大数据人才，满足即将爆发的企业的大数据人才需求。

自2013年起，中国计算机学会就已经开始面向高校师生及社会在职人员开展全国大数据技术创新大赛，至今已经持续四届。自2015年国家提出大数据战略后，各类大数据技术竞赛可谓是层出不穷，2016年仅半年时间就有多个赛项在全国各地开展，包括“中国高校计算机大赛——大数据挑战赛”、“优易数据杯”全国开放大数据创新创意大赛、“上海BOT大数据应用大赛”等。

2017年全国职业院校技能大赛“大数据技术与应用”赛项（高职组）在贵州轻工职业技术学院举行，全国28个省市区、78支参赛队伍，共计近500余人参与了比赛。

组织该竞赛的目的主要有三个。第一，通过竞赛提高学生运用大数据技术解决实际问题的能力，促进高职院校大数据相关专业教师和学生的技术技能提升。第二，通过在竞赛中引入企业应用的最新、最普及的技术，促进高职院校大数据相关专业发展，提升高职院校大数据专业对接产业对接行业的能力。第三，通过竞赛建立完整的大数据专业人才培养体系，通过大赛资源转化，建设相关专业配套教学资源，让大赛服务教学、服务国家人才培养战略，发掘和培养更多满足大数据市场需求的优秀技术技能型人才。

## **赛项设计原则**

## 公开、公平、公正；

坚持公平、公正、公开。赛项方案的设计、赛项过程的组织、赛项结果的评判都本着公平、公正、公开的原则，由赛项专家组、赛项裁判组提前制定详细的赛项组织流程、赛项评判细则，整个竞赛过程透明公开，实时发布竞赛进程和成绩。

## 常规赛项关联专业人才需求量大，服务国家重点战略。

赛项各环节涉及的岗位较多，涵盖了数据开发工程师、大数据分析师、大数据架构师、大数据系统工程师等多个岗位方向，大数据技术的应用已经不仅在互联网行业，各行各业都已经在使用大数据技术提升效率。

2017将会有越来越多的公司加速采用大数据技术。互联网数据中心(IDC)预测，到2020年大数据和分析技术市场，将从今年的1301亿美元增加至2030亿美元。大数据产业将聚焦于开源组织和开放源码的基础上，从内存技术、机器学习、预测分析、智能化终端、物联网、边缘计算等核心技术与应用领域相结合，推动社会各行业大数据的共享与获取数据价值手段的更新与变革。

Gartner公司预测大数据将为全球带来440万个IT新岗位和上千万个非IT岗位，Gartner公司预测在未来6年，仅在美国本土就可能面临缺乏 14万至19万具备深入分析数据能力人才的情况，同时具备通过分析大数据并为企业做出有效决策的数据的管理人员和分析师也有150万人的缺口。

中国政府已经把大数据作为国家战略来对待，并列入十三五计划。大数据正在成为经济社会发展新的驱动力。在大数据时代，各行业都会产生大量数据，所以每个行业对大数据人才都有着大量的需求。培养大数据人才将成为所有高校的当务之急。

## 竞赛内容对应相关职业岗位或岗位群、体现专业核心能力与核心知识、涵盖丰富的专业知识与专业技能点；

经过与众多的互联网大企业及中小型大数据服务型企业调研，大数据岗位群主要分为：大数据算法开发方向、大数据应用平台开发方向、大数据应用方向、大数据分析方向、大数据系统方向等五个方向，而适合高职层次的岗位主要包含三大类岗位群方向：大数据应用方向、大数据分析方向、大数据系统方向。

**大数据应用方向：**主要包含了大数据系统研发工程师、大数据应用开发工程师等岗位，需要具备掌握大数据的基本概念和应用领域，掌握程序设计和实现方法，掌握使用HBase、HDFS存储和管理数据的方法，具有开发Java Web应用的能力，具有基于MapReduce开发大数据应用开发的能力；

**大数据分析方向：**主要包含了大数据分析工程师、大数据可视化工程师等岗位，需要掌握大数据的基本概念和应用领域，掌握程序设计和实现方法，掌握使用HBase、HDFS存储和管理数据的方法，具有使用Mahout数据挖掘算法进行数据分析的能力，具有基于MapReduce开发大数据应用开发的能力；

**大数据系统方向：**主要包含了大数据系统运维工程师、大数据系统搭建工程师的岗位，掌握大数据的基本概念和应用领域，掌握大数据系统的架构和设计方法，掌握计算机网络基本原理和基本知识，熟练掌握Linux Shell的使用，熟悉Linux操作系统，具有大数据系统的搭建、调优和运维能力。

针对产业面向的岗位和岗位群以及人才的培养标准，在本赛项中重点体现了岗位群中关于大数据采集、大数据清洗、大数据分析，大数据可视化、系统搭建等部分内容，涵盖了专业核心知识，满足产业、岗位对人才的技术技能要求。

## 竞赛平台成熟。根据行业特点，赛项选择相对先进、通用性强、社会保有量高的开源技术与开源软件。

根据行业特点，赛项选择相对先进、通用性强、社会保有量高的设备与软件。在设计赛项竞赛内容时，综合考虑对应的技术平台的可行性、稳定性以及院校的保有性等因素，大赛需要采用的业界最为主流的开源技术与开源软件，硬件设备中使用的PC、服务器和交换机几乎全部职业院校的实验室早已经广泛使用。大赛所使用的竞赛平台已在各省院校得到广泛使用，且在各地的省赛和行业赛中得到成功应用，并获得普遍好评。

## **赛项方案的特色与创新点**

本赛项方案在如下四个方面体现了特色和创新点

## 紧密结合大数据产业需求

大数据行业发展迅猛，急需大量大数据专业人才（既有较强的操作能力、又有一定的大数据技术理论知识），以满足大数据的环境部署与运维、大数据数据挖掘与应用开发、大数据可视化等专业岗位的人才需求。本赛项以巩固学生的大数据专业理论知识为前提，强化学生大数据实践操作能力为目的，培养大数据应用技术技能型专业人才。本次赛项设计顺应大数据产业经济发展趋势，在赛项内容设计上，紧密联系行业，有效对接大数据产业人才发展及技能需求，提高选手大数据环境运维、大数据分析与应用、大数据可视化开发能力，满足各行业大数据应用发展对人才培养的需求。

## 再现真实岗位工作流程

本次赛项竞赛项目源自真实大数据电子商务应用商业项目，涉及大数据运维工程师、大数据分析与挖掘算法应用工程师、大数据可视化工程师重要工作岗位，完整再现大数据运维工程师、大数据分析与挖掘算法应用工程师、大数据可视化工程师真实岗位工作流程，通过竞赛训练，使学生掌握大数据环境部署与运维、大数据数据挖掘算法应用、大数据可视化开发全过程技能，从而达到理论与实践的紧密结合，教学标准与行业标准的高度统一。

## 专业能力+创新能力的科学量化评分体系

本次赛项重点比拼参赛选手在大数据领域各个业务流程及岗位中所具备的综合技能,涵盖大数据基础环境运维、数据挖掘算法应用、大数据可视化专业技能、应对海量数据存储与管理、大数据算法领会与应用的创新能力，评分过程科学且可量化，既有大数据基础环境Hadoop环境运维的考核，又有对数据挖掘算法与应用的考核，还有大数据可视化呈现能力的考核，同时，竞赛设计融合理论知识，应用数据挖掘算法，结合实战项目应用大数据技术，构建“专业能力+创新能力”的科学量化评分体系。

## 推动大数据技术的转化应用

赛项的主要内容包括：如何熟练使用大数据的工具，还能熟练掌握数据挖掘是如何对数据清洗、建模、参数调优、可视化结果输出。而这些内容又是行业大数据应用最核心的环节，通过举行比赛，加快大数据技术的转化与应用，更好更快地推动大数据技术在各行业的发展。同时通过大赛能检验学生对各大数据技术生态核心组件基础理论和应用等方面的技能水平，也检验了大数据专业群的建设水准。通过竞赛这个交流、合作、竞技、发展的平台，较好地引导和推动了专业的建设和可持续发展。

## **竞赛内容简介**

赛项名称：大数据处理与分析（Big Data Processing And Analysis）

赛项简介：

赛项以大数据为核心内容和工作基础，在互联网、云计算、大数据应用的产业背景下，根据给定的应用项目需求，完成大数据平台的软件安装部署；根据任务的描述，完成大数据应用程序的开发，形成正确的大数据应用分析结果，最后提交每项任务的报告。

竞赛的主要目的是考察学生应用大数据知识解决实际问题的能力。参赛选手需要针对竞赛选题，设计构建业务逻辑模型，形成业务设计文档，考察选手的业务分析能力；基于大数据实训平台，综合运用数据抽取工具Sqoop/Flume/Kafka、分布式存储系统HDFS、分布式计算框架MapReduce/Yarn、内存计算Spark、流式计算Storm、数据挖掘方法库Mahout、数据仓库Hive、深度学习等工具和技术，匹配和连接数据源，完成大数据的提取、清洗、转换、分析、挖掘操作，产生分析结果；依据项目应用需求和分析结果，完成标准化的大数据分析报告。

本赛项是对参赛选手大数据知识和应用能力的综合检验，在一定程度上促进了高职院校相关专业及课程建设，促进学生综合应用能力和职业素质的同步提高，提升毕业生的就业竞争力。

The competition use big data as the core content and basis for the work, in the industry background of Internet, cloud computing, big data applications.The players will install and deployment big data platform according to the requirement of the given application project,and write programs to solve task problems, and finally submit the report of each task.

The main purpose of the contest is to investigate the ability of students to solve practical problems with big data knowledge.The players need to design business architecture, logic model and form the business design documentation based on the competition topics, investigate players’ business analytical skills; match and link data sources, complete big data extraction, cleaning, transformation, analysis, mining operations, generate analysis results in the integrated use of data extraction tools Sqoop / Flume / Kafka, a distributed storage system HDFS, distributed computing framework MapReduce / Yarn, memory computing Spark, stream computing Storm, data mining library Mahout, data warehouse Hive, the depth of learning tools and techniques based on big data training platform; according to the project application requirements and analysis, complete standardization of big data analysis reports.

This competition is a integrated test in the Big Data knowledge and ability of the players, to a certain extent, it promoted the Higher Vocational Colleges’ relevant professional and curriculum development, promoted students’ comprehensive ability and professionalism improve synchronization, enhanced the competitiveness of graduates employment.

## **竞赛方式（含组队要求、是否邀请境外代表队参赛）**

（一）本赛项为团体赛，每支参赛队由3名选手组成,须为同校在籍高职学生，其中队长1名，性别和年级不限，最多2名指导教师。不计选手个人成绩，统计竞赛队的总成绩进行排序。

（二）本赛项不邀请国际团队参赛，但将推荐优秀赛队参加相应的国际赛事。

（三）本赛项设单一场次，所有参赛队在现场根据给定的项目任务，在4小时内相互配合，在设备上完成比赛内容，最后以设备代码文件、提交的截图、文档和竞赛作品作为最终评分依据。

## **竞赛时间安排与流程**

比赛时限：4小时

详细安排及流程

|  |
| --- |
| **日程安排** |
| 07:00-07:10 | 裁判进入裁判室 |
| 07:10-07:50 | 选手抽签，完成一次和二次加密，入场 |
| 07:50-08:00 | 参赛代表队就位并领取比赛任务 |
| 08:00-12:00 | 比赛时间 |
| 12:00-12:10 | 参赛代表队离场 |
| 12:00—14:00 | 仲裁受理 |
| 12:10—13:10 | 进行三次加密 |
| 12:30—13:30 | 裁判培训 |
| 13:30-20:00 | 裁判评分 |

**各参赛队报到**

**赛前领队会**

**领队参观场地**

**参赛队赛场检录**

**赛队抽签和二次加密**

**参赛队确认设备**

**参赛队竞赛**

**裁判加密评分核分**

**闭赛式颁奖**

**赛前**

**比赛**

**申诉受理**

**是否有申诉诉**

**是**

**否**

## **竞赛试题**

根据全国职业院校技能大赛要求以及适合职业院校学生掌握的大数据应用技术的知识体系情况，设置本赛项竞赛试题，并承诺保证于开赛1个月前在大赛网络信息发布平台上（www.chinaskills-jsw.org)公开全部赛题。竞赛试题的内容形式等可参见附件一的竞赛样题。

## **评分标准制定原则、评分方法、评分细则**

## 评分标准制定原则

1. 竞赛评分严格按照公平、公正、公开、科学、规范的原则。
2. 相关文档的准确性与规范性；
3. 团队风貌、团队协作与沟通、组织与管理能力；

## 评分方法

1.竞赛采取三次加密。第一组加密裁判组织参赛队选手第一次抽签，抽取参赛编号，替代选手参赛证等个人信息；第二组加密裁判组织参赛选手进行第二次抽签，确定赛位号，替换选手参赛编号；第三组加密裁判对各参赛队竞赛结果进行加密，替换赛位号。三次加密信息由不同加密裁判密封后保管，在评分结束后开封统计成绩。

2.竞赛对参赛队伍提交的作品采取客观性结果评分。采取分步得分、累计总分的计分方式。各环节分别计算得分，错误不传递，按规定比例计入团队总分。根据赛题情况划分模块，每两名裁判负责一个模块进行独立评分，取两名评分裁判一致认可评分作为该参赛队该模块的最后得分。两名裁判评分差距超过5%时，由裁判长主持复核。裁判长在竞赛结束18小时内提交评分结果，经复核无误，由裁判长、监督人员和仲裁人员签字确认后公布。

3.裁判长正式提交评分结果并复核无误后，加密裁判在监督人员监督下进行三次解密：竞赛作品编号到工位号解密；工位号到参赛编号解密；参赛编号到参赛队名称解密。

4.为保障成绩评判的准确性，监督组对赛项总成绩排名前30%的所有参赛队伍的成绩进行复核；其余成绩进行抽检复核，抽检覆盖率不低于15%。

5.监督组在复检中发现错误，需以书面形式及时告知裁判长，由裁判长更正成绩并签字确认。如复核、抽检错误率超过5%，裁判组需对所有成绩进行复核。

6.在竞赛过程中，参赛选手如有不服从裁判裁决、扰乱赛场秩序、舞弊等行为的，由裁判长按照规定扣减相应分数，情节严重的将取消比赛资格，比赛成绩计0分。

## 评分细则

考察参赛团队的系统设计和描述能力。每项从主要考合理性、完整性、清晰性方面进行考核。

|  |  |  |
| --- | --- | --- |
| **考核环节** | **评分维度** | **分值** |
| 大数据环境的搭建（180分） | 任务1 Hadoop安装与部署 | 60 |
| 任务2 Hive安装配置 | 60 |
| 任务3 Spark的安装配置 | 60 |
| 大数据应用（120分） | 任务1 MapReduce统计计算 | 60 |
| 任务2 Hadoop数据压缩与解压 | 60 |
| 数据抓取(100分) | 任务1 数据抓取 | 100 |
| 数据分析与展现（200分） | 任务1 Hive导入、查询统计 | 100 |
| 任务2 数据可视化编程 | 100 |
| 数据挖掘与分析（400分） | 任务1 销量预测 | 200 |
| 任务2 缺失值处理 | 200 |

## **奖项设置**

本赛项奖项设团体奖。设奖比例为：以赛项实际参赛队总数为基数，一、二、三等奖获奖比例分别为10%、20%、30%（小数点后四舍五入）。

获得一等奖的参赛队指导教师获“优秀指导教师奖”。

## **技术规范**

|  |  |  |
| --- | --- | --- |
| **序号** | **标准号** | **中文标准名称** |
| 1 | GB/T 18142-2000 | 信息技术数据元素值格式记法 |
| 2 | GB/T 18391.1-2009 | 信息技术元数据注册系统(MDR)第1部分：框架 |
| 3 | GB/T 18391.2-2009 | 信息技术元数据注册系统(MDR)第2部分：分类 |
| 4 | GB/T 18391.3-2009 | 信息技术元数据注册系统(MDR)第3部分：注册系统元模型与基本属性 |
| 5 | GB/T 18391.4-2009 | 信息技术元数据注册系统(MDR)第4部分：数据定义的形成 |
| 6 | GB/T 18391.5-2009 | 信息技术元数据注册系统(MDR)第5部分：命名和标识原则 |
| 7 | GB/T 18391.6-2009 | 信息技术元数据注册系统(MDR)第6部分：注册 |
| 8 | GB/T 21025-2007 | XML使用指南 |
| 9 | GB/T 23824.1-2009 | 信息技术实现元数据注册系统内容一致性的规程第1 部分：数据元 |
| 10 | GB/T 23824.3-2009 | 信息技术实现元数据注册系统内容一致性的规程第3 部分：值域 |
| 11 | GB/T 20009-2005 | 信息安全技术数据库管理系统安全评估准则 |
| 12 | GB/T 20273-2006 | 信息安全技术数据库管理系统安全技术要求 |
| 13 | GB/T 22080-2008 | 信息技术安全技术信息安全管理体系要求 |
| 14 | GB/T 22081-2008 | 信息技术安全技术信息安全管理实用规则 |
| 15 | 20100383-T-469 | 信息技术安全技术信息安全管理体系实施指南 |
| 16 | GB/T 29262-2012 | 信息技术面向服务的体系结构（SOA）术语 |
| 17 | GB/T 29263-2012 | 信息技术面向服务的体系结构（SOA）应用的总体技术要求 |
| 18 | DG/TJ08-2068-2009 | 公共建筑用能分项计量系统工程技术规范 |

## **建议使用的比赛器材、技术平台和场地要求**

需要特殊提到的是：本赛项的竞赛选题基于人工智能主流技术设计完成，不依赖于特定平台。学校在搭建赛前训练模拟环境时，除可选用上述（详见本文中第四章节第4点）人工智能平台外，也可选择其他人工智能平台产品（采用相似技术）完成。

**技术平台：**

要求至少满足 90 支参数队伍 270 人规模。每支队伍独自一套环境作为调试测试。

1.集群配置清单

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **设备名称** | **设备基本功能描述** | **数量** |
| 1 | 云环境节点服务器 | intel E5-2630v3×2/256GB内存/1TB×3 7200转 SATA 6Gb 硬盘/1TB×2 SAS盘/支持RAID5/1Gb ×1管理接口/1Gb ×2业务网络接口/10Gb ×2数据网络接口 | 23台 |
| 2 | 大数据教学实训平台 | 平台负责整个产品云平台中的计算资源、存储资源、和网络资源的虚拟化实现，并提供分配、共享、管理、优化、和运维等功能，包含：1、采用目前业界先进的互联网和云计算设计架构思想、虚拟化技术和云管理平台技术实现；2、整体平台提供统一的虚拟化资源池和物理资源池的运行环境和管理功能。3、提供大数据集群环境部署，并能满足实训需求。 | 1套 |
| 3 | 万兆交换机 | 48口全万兆交换机，用于云环境控制节点服务器与计算节点服务器之间分布式存储网络的互联。 | 1台 |
| 4 | 千兆交换机 | 48口全千兆交换机，用于云环境控制节点服务器与计算节点服务器之间管理网络的互联，以及所有PC与节点服务器的互联。 | 2台 |

2.每支队伍的设备配置

|  |  |  |  |
| --- | --- | --- | --- |
| **设备类别** | **数量** | **设备用途** | **基本配置** |
| 大数据集群服务器 | 3台 虚拟机，其中一台作为Hadoop/Spark为主节点，另外2台作为从节点。 | 内置Hadoop、Spark大数据环境，支撑运行数据挖掘、机器学习、深度学习模型训练。 | CPU主频2.00GHz以上, 双核以上；内存6G以上；SATA硬盘 80 G 以上, 7200 转；1000M 以上网络接口。 |
| 客户端 | 每支参赛队伍3台。根据参赛团队数量，配备30%的备份机器。 | 竞赛选手比赛使用。 | 2.0GHZ以上处理器，4G以上内存，显示器要求1024\*768以上 |
| 大数据教学实训平台 | 每支参赛队伍1组帐号，3个子帐号。 | 大数据教学实训平台是基于open-stack基础云平台搭建，提供统一的虚拟化资源池和物理资源池的运行环境和管理功能，满足教学实训需求。 |  |

3.软件环境

|  |  |  |
| --- | --- | --- |
| **设备类别** | **软件名称** | **版本配置** |
| 服务器 | 操作系统 | CentOS 7 64 位 |
| Java | JDK 1.8 |
| 百度AI SDK | https://ai.baidu.com/资源 |
| AI实训/比赛云平台 | 版本 V1.0 |
| Hadoop | Hadoop版本2.7 |
| Spark | Spark 版本 2.1.0 |
| Scala | Scala 版本2.11 |
| Python | 版本 2.7已安装的库:Numpy: 数组、矩阵运算库matplotlib ：图形绘制库Scrapy：爬虫库 |
| 客户端 | 操作系统 | Windows 7 64 位 |
| Java | JDK 1.8 |
| Eclipse | Eclipse jee oxygen |
| Maven | Maven 3.0 以上 |
| Python | 版本 2.7已安装的库:Numpy: 数组、矩阵运算库matplotlib ：图形绘制库Scrapy：爬虫库 |
| Scala | 版本2.11 |
| 浏览器 | Chrome |
| 文档编辑器 | Office 2007 以上版本 |

## 竞赛环境场地要求

竞赛场地包括选手竞赛区、裁判区、观摩展示区和安全救助区。环境标准要求保证赛场采光(大于500lux)、照明和通风良好。

### 1.选手竞赛区

选手竞赛区域的每间工位内设有操作平台并配备220伏电源，工位内的电缆线应符合安全要求。每间竞赛工位面积9～10㎡，参赛队之间采用挡板进行隔离确保互不干扰。竞赛工位标明工位号，并配备竞赛平台和技术工作要求的软硬件。

### 2.裁判区域

设置指定的裁判工作场地，裁判区域配置工作PC，裁判场地满足工作需要。

### 3.观摩展示区

观摩展示区与选手竞赛区域隔离，配备大屏幕和计时装置，展示赛场竞赛效果，监控比赛进度，体现大赛“公正、公开、公平”的竞赛原则。

### 4.安全救助区

设置安全救助区，应对参赛选手的突发事件，提供医疗救护设施和专业人员，确保比赛正常进行和选手安全。

## **安全保障**

(一) 成立安全保障工作组，负责本赛项筹备和比赛期间的各项安全工作，赛项执委会主任和安全保障工作组组长均为第一责任人；

(二) 制定安全管理的相应规范、流程和突发事件应急预案，保证比赛筹备和实施工作全过程的安全：

1.比赛期间发生意外事故，发现者应第一时间报告赛项执委会，同时采取措施避免事态扩大。

2.赛项执委会应立即启动预案予以解决并报告赛区执委会。赛项出现重大安全问题可以停赛，是否停赛由赛区组委会决定。

3.事后，赛区执委会应向大赛执委会报告详细情况。

(三) 确保比赛内容涉及的器材、设备符合国家有关安全规定；

(四) 赛项执委会在赛前对本赛项全体裁判员、工作人员、指导老师、参赛选手进行安全培训；

(五) 赛项执委会制定专门方案保证比赛命题、赛题保管、发放、回收和评判过程的安全；

(六) 赛项执委会在赛前组织专人对比赛现场、住宿场所和交通保障进行考察，并对安全工作提出明确要求。赛场的布置，赛场内的器材、设备，应符合国家有关安全规定，组织模拟测试，排除安全隐患；

(七) 赛场周围设立警戒线，防止无关人员进入，发生意外事件。比赛现场内参照相关职业岗位的要求为选手提供必要的劳动保护和医务服务；

(八) 承办院校应提供保障应急预案实施的条件，明确制度和预案，并配备急救人员与设施；

(九) 赛项执委会会同承办院校制定开放赛场和体验区的人员疏导方案。赛场环境中存在人员密集、车流人流交错的区域，除了设置齐全的指示标志外，增加引导人员，并开辟备用通道；

(十) 大赛期间，赛项承办院校须在赛场设置医疗医护工作站；在管理的关键岗位，增加力量，建立安全管理日志。比赛期间安排的住宿地应具有宾馆、住宿经营许可资质，保证住宿、卫生、饮食安全等；

(十一) 比赛期间发生意外事故时，发现者应第一时间报告赛项执委会，同时采取措施，避免事态扩大。赛项执委会应立即启动预案予以解决并向赛区执委会报告。

## **经费概算**

参照《全国职业院校技能大赛经费管理暂行办法》的相关要求，制定本赛项经费概算。

赛项专用保障经费预算人民币陆拾万元整，采用企业统筹方式提供。预算支出支出项目如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **费用类别** | **项目** | **金额****（万元）** | **备注** |
| 赛项研讨论证、赛题开发及培训预算 | 竞赛方案研讨论证会议组织 | 5 | 研讨会专家的用餐、论证费等 |
| 竞赛试题开发 | 10 | 专家封闭开发住宿、用餐等 |
| 印刷费 | 2 | 大赛通知、赛程、竞赛文档、赛题等 |
| 培训费 | 1 | 裁判员培训费用 |
| 专家 | 3 | 包括命题、评审、裁判等环节的专家费用 |
| 赛项组织预算 | 场地布置 | 8 | 根据比赛要求布置场地、布线、主席台搭建等 |
| 设备运输、安装、调试 | 6 | 比赛设备运输及安装等 |
| 现场技术支持 | 3 | 竞赛现场支持 |
| 赛项宣传费用 | 5 | 赛场环境设计与制作、宣传材料制作，平面媒体、网站宣传维护等 |
| 服装费 | 4 | 选手、指导老师、裁判、监考、工作人员、志愿服装费 |
| 获奖奖励 | 4 | 获奖人员奖励、奖品和纪念品等 |
| 餐费、交通费 | 6 | 领导、专家、选手在津食、交通费用、工作人员餐费 |
| 预留资金 | 3 | 处理赛事突发事件 |
| 总计 | 60 |  |

## **比赛组织与管理**

将成立赛项执行委员会，下设专家工作组、组织保障工作组和安全保障工作组。经大赛执委会核准发文后成立。

赛项执委会：全面负责本赛项的筹备与实施工作，接受大赛执委会领导，接受赛项所在分赛区执委会的协调和指导。赛项执委会的主要职责包括：领导、协调赛项专家组和赛项承办院校开展本赛项的组织工作，管理赛项经费，选荐赛项专家组人员及赛项裁判与仲裁人员等；

赛项专家组：在赛项执委会领导下开展工作，负责本赛项技术文件编撰、赛题设计、赛场设计、设备拟定、赛事咨询、技术评点、赛事成果转化、赛项裁判人员培训、赛项说明会组织等竞赛技术工作；同时负责赛项展示体验及宣传方案设计；

承办院校：在赛项执委会领导下，负责承办赛项的具体保障实施工作，主要职责包括：按照赛项技术方案要求落实比赛场地及基础设施，赛项宣传，组织开展各项赛期活动，参赛人员接待，生活服务，比赛过程文件存档等工作，赛务人员及服务志愿者的组织，赛场秩序维持及安全保障，赛后搜集整理大赛影像文字资料上报大赛执委会等。赛项承办院校按照赛项预算执行各项支出。承办院校人员不得参与所承办赛项的赛题设计和裁判工作；

现场裁判、仲裁、监督组：开赛前一周，在裁判员库、仲裁员库、监督员库中随机抽取组成。裁判组负责赛前检查及赛场鉴定、现场执裁和评审比赛结果等工作；仲裁组负责受理各参赛队的书面申诉、对受理的申诉进行深入调查，做出客观、公正的集体仲裁；监督组对指定赛区、赛项执委会的竞赛筹备与组织工作实施全程现场监督，包括赛项竞赛场地和设施的部署、选手抽签、裁判培训、竞赛组织、成绩评判及汇总、成绩发布、申诉仲裁、成绩复核等；

协办企业：提供竞赛现场设备并设置技术保障组，为竞赛设备、软件与竞赛设施提供保养、维修等服务，保障设备的完好性和正常使用，保障设备配件与操作工具的及时供应。

## **教学资源转化建设方案**

参照《2017年全国职业院校技能大赛赛项资源转化工作办法》，制定本赛项资源转化方案。

|  |  |  |  |
| --- | --- | --- | --- |
| **内容** | **要求** | **完成时间** | **负责单位** |
| 赛项风采展示 | 画面精美、伴音动听、播放流畅，时间长度15分钟左右的赛项宣传片及获奖代表优秀选手、指导教师风采展示片；竞赛过程音视频记录。 | 赛项闭赛后5日内 | 承办院校 |
| 点评材料 | 评委、裁判、专家点评 | 赛项闭赛后5日内 | 赛项执委会 |
| 竞赛试题库及技术分析报告 | 10套试题、竞赛技能考核评分案例及技术分析报告 | 赛项闭赛后30日内 | 赛项执委会 |
| 大数据技术专业课程建设和教学实验平台建设方案指导书 | 结合当前各行业推动大数据技术发展，以及对人才岗位能力需求的变化，对大数据技术专业人才培养方案中课程建设和教学实验平台建设方案进行优化，并输出指导书。 | 赛项闭赛后100日内 | 赛项执委会 |
| 课程资源库 | 根据大数据技术与应用专业的课程体系，结合赛题考试知识技能点以及大数据教学相关要求，开发系列课程，并配备丰富的课件、教案、训练指导等教学资源。 | 赛项闭赛后100日内 | 赛项执委会 |
| 高职大数据技术专业教师引导培养和能力评估方案 | 根据高职大数据新专业的建立和人才培养目标，建设一套针对高职大数据专业教师培养方案和能力评估方案。 | 赛项闭赛100日内 | 赛项执委会 |

## **筹备工作进度时间表**

依据赛项筹备工作，制定筹备工作时间进度表。

| **序号** | **事项及内容** | **开始时间** |
| --- | --- | --- |
| 1 | 赛项申报 | 2017年9月上旬 |
| 2 | 确认赛项执委会 | 2017年10月上旬 |
| 3 | 专家组筹备会议 | 2017年10月下旬 |
| 4 | 专家组第一次会议 | 2017年10月下旬 |
| 5 | 确定比赛场地 | 2018年1月上旬 |
| 6 | 赛事设备测试 | 2018年1月中旬 |
| 7 | 专家组第二次会议，确定样题和评分细则 | 2018年3月上旬 |
| 8 | 参赛队报名 | 2018年3月中旬 |
| 9 | 召开赛项说明会 | 2018年3月下旬 |
| 10 | 服装数量规格确定 | 2018年4月上旬 |
| 11 | 场地布置规划 | 2018年4月上旬 |
| 12 | 成立命题专家组 | 2018年4月上旬 |
| 13 | 服装制作 | 2018年4月上旬 |
| 14 | 确定裁判名单 | 2018年4月中旬 |
| 15 | 竞赛指南 | 2018年4月下旬 |
| 16 | 赛场设备安装调试 | 2018年4月下旬 |
| 17 | 专家封闭命题 | 2018年4月下旬 |
| 18 | 专家、裁判培训 | 2018年5月上旬 |
| 19 | 召开赛前领队会 | 2018年5月中旬 |
| 20 | 开幕式 | 2018年5月中旬 |
| 21 | 正式比赛 | 2018年5月中旬 |
| 22 | 闭幕式 | 2018年5月下旬 |

## **裁判人员建议**

按照《2017年全国职业院校技能大赛专家和裁判工作管理办法》要求，裁判由行业、企业对职业教育有理解和认识的专家、在职业教育教学改革上有成果的职业院校教师组成。裁判组设立裁判长一名。根据大赛工作需要，裁判分为加密裁判、现场裁判和评分裁判三类。

加密裁判。负责组织参赛队伍（选手）抽签并对参赛队伍（选手）的信息进行加密、解密。

现场裁判。按规定做好赛场记录，维护赛场纪律，对参赛队伍（选手）的操作规范、现场环境安全等进行评定；

评分裁判。负责对参赛队伍（选手）的技能展示、操作规范和竞赛作品等按赛项评分标准进行评定。

裁判数量共计30名,要求：

1.具有良好的职业道德和心理素质，严守竞赛纪律，服从组织安排，责任心强；

2.现场裁判和评分裁判须从事赛项所涉及专业（职业）相关工作5年以上（含5年），具备深厚的专业理论知识和较高的实践技能水平，熟悉职业教育和大赛工作，原则上应具有中级及以上专业技术职称或高级技师职业资格，有省级或以上职业技能竞赛执裁经验者优先考虑；

3.有较强的组织协调能力和临场应变能力；

4.年龄原则上不超过65周岁，身体健康，无任何违法违纪记录，且获得工作单位支持，能在规定时间内到岗，并按要求完成指定裁判工作。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **裁判类别** | **知识能力要求** | **专业技术职称****（职业资格等级）** | **人数** |
| 1 | 裁判长 | 计算机网络、计算机应用 | 高级职称以上或企业高级工程师 | 1 |
| 2 | 加密裁判 | 无 | 无 | 3 |
| 3 | 现场裁判 | 计算机相关 | 无 | 4 |
| 4 | 评分裁判 | 云计算、大数据方向 | 中级职称以上的教师或企业的工程师 | 9 |
| 5 | 计算机网络方向 | 中级职称以上的教师或企业的工程师 | 6 |
| 6 | 软件开发方向 | 中级职称以上的教师或企业的工程师 | 7 |
| 裁判总人数 |  |  |  | 30 |

## **其他**

为办好本次大赛将免费提供培训、赛前说明会。建立赛项在线学习平台、QQ技术服务群、400客服电话、在线问答系统等沟通渠道，推送培训信息和资料，同老师进行互动、解答疑问，做好服务工作，推动大数据专业的快速发展。

**（一）赛项联络人员**

## **附件一：**

大数据处理与分析大赛任务书（样题）

## **题目：大数据在电商场景的应用**

## 赛题背景

大数据在电商场景下的应用十分广泛，用户画像、精准营销、中文图谱以及预测分析等应用取得了十分明显的效果。某电商公司已上线运营平台多年，数据的使用只停留在初级统计分析，为提高数据的使用价值，通过大数据的分析方法帮助企业提高销售和促销水平，为企业决策提供更多元且有价值的支持信息，公司计划聘请“云购商城平台”大数据分析公司，收集并分析该商城用户访问日志和订单信息。合作之前需提出可行的技术解决方案。

解题本次竞赛，使用业界广泛的“Python”语言作为开发分析程序的基础语言，并综合利用Hadopp、MapReduce、Hive、Spark等技术提高开发效率，所有工作集中在利用Hadoop平台数据处理能力，并利用Hive等技术提高效能和简化MapReduce过程，采用Spark对数据分类，回归，聚类，得到相关数据模型，也可以完成解答本次试题。

## 项目任务

### **第一部分 大数据环境搭建**

1. Hadoop安装与部署,按下面步骤在指定centos 7.0系统下部署Hadoop伪分布式平台，并根据计算对象调优Hadoop平台的性能（ centos 系统已完成各种升级和JDK环境的安装）。（60分）
	* 1. 搭建Hadoop 环境，要求Hadoop安装在/opt/modulse/hadoop这个目录下。
		2. 环境配置成功后， 格式化 HDFS 文件系统，启动HDFS,启动YARN。执行 jps 命令，把jps结果截图保存为 result\_0101.jpg 文件。
		3. 配置主机名为hadoop-host , 在hdfs启动成功后，用命令在hdfs文件系统上创建文件夹hadoop-hdfs ，并在hadoop-hdfs文件夹下创建hdfs.txt，创建成功后，在本地浏览器中访问hdfs文件系统：<http://hadoop-host:50070> ，在文件系统里找到刚创建的hdfs.txt文件，并截屏保存为result\_0102.jpg文件。
2. 完成Hive环境搭建。搭建完成后，启动Hive 外壳环境(shell)， 执行查看表的命令，将执行结果截图保存为 result\_0103.jpg。 （60分）
3. 完成Spark环境搭建。搭建完成后，启动 spark shell，截图保存为result\_0104.jpg。（60分）

### **第二部分 大数据应用**

1. MapReduce统计计算,按下面步骤建立用户目录，并导入存于竞赛平台data目录中数据文件file0201.dat（file0201.dat为赛前抽取的数据文件）。（60分）
2. 创建input目录，把数据文件上传至input目录。
3. 使用MapReduce统计file0201.dat 文件中每个以空格分割的单词出现的次数。
4. 将每个单词的统计结果以“单词 空格 次数”格式按行输出，按单词字母排序顺序排序。列如:

my 10

say 3

本题要求将第3步运行结果保存为result\_0201.txt文件。

1. 在Hadoop 平台进行数据压缩和解压，并导入平台data目录中数据文件file0202.dat，对它先压缩，在解压，并读取解压后的数据文件。（60分）

本题要求将压缩后与解压后的文件截图保存为result\_0202.jpg。

本题要求将读取解压后的文件数据截图保存为result\_0203.jpg。

### **第三部分 数据抓取**

1. 数据抓取（100分）

网页 http://\*\*\*/goodsid=?\*\*\* （\*\*\*代表的具体访问路径待定）中包含用户对某件商品的的评价信息，请编写程序抓取网页上商品的评价数据, 并按下面字段的顺序保存成csv文件，命名为comment\_spider.csv保存。（100分）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户id | 商品id | 评分 | 评价内容文本 | 商品类型 | 评论时间 | 购买时间 |

### **第四部分 数据分析与展现**

1. Hive导入、查询统计（100分）

本题给定的数据为 order.csv , 规模约1G。其中 order.csv 的字段顺序如下， 后面还会用到该文件，字段顺序都一样，不再赘述。

userId,goodsId,shopId,createTime,price,firstCategory,secondCategory,thirdCategory

编写Hive命令建立数据表，并将order.csv 导入所建立的数据表，然后编写Hive查询语句获取2016全年销量最高的10件商品记录，并将查询结果输出到top\_product.txt。以“商品id, 销量” 格式按行输出，按销量从高到低排序，列如

100010001 104059

100010002 8832

1. 数据可视化编程（100分）

本题给定的数据为 order.csv，规模约1G，对店铺A中的商品B 统计每天的销量，编程画出某月的销量折线图（可以使用 matplotlib ），其中X轴为时间，只取天数，如 01，Y 轴为销量。图片保存为 result\_daily\_order.jpg。

### **第五部分 数据挖掘与分析**

采用Spark算法库（MLlib）提供的算法完成下面需求

1. 销量预测

order.csv中包含了来自不同店铺不同商品的订单信息，先对order.csv中的店铺A中的商品B的销量进行统计。然后建立商品价格和商品销量模型。销售人员对下一个月商品B制定了3种不同的价格，现在销售人员想知道下一个月3种不同价格分别对应的总销量, 以便确定下一个月商品最终的价格。请您给出预测。具体任务说明如下：

1. 数据预处理，从原始数据中提取店铺A中的商品B的销量统计，按（价格、每日平均销量）进行统计。输出price\_sale.csv 文件，格式如下：（50分）

价格1, 每日平均销量

价格2, 每日平均销量

1. 选择算法，建立模型，输出建模word文档,命名为 price\_sale\_modle.doc 。（50分）
2. 编写代码，对预处理好的数据进行训练，得到模型文件。为了符合开放性和创造性，解题思路可以不遵循固定的算法，不一定要输出模型，因为有些算法（比如KNN）没有模型输出。（80分）
3. 对3种不同价格分别预测每日平均销量，一个月按30天计算，得出下一个月的总销量。预测的结果按下面的csv 格式保存，文件命名为sale\_predict.csv 保存。（20分）

价格1, 1000

价格2, 1200

价格3, 800

1. 缺失值处理：补充用户性别。

有些注册用户的某些基础信息是不完善的，比如性别、年龄等等。系统知道用户的信息越多，那么系统推荐的商品将越精准。这不论对电商还是用户都是有利的。我们可以根据用户的购买行为，大致判断用户的性别。本题给定的数据为 order.csv 和 user.csv 。user.csv 中有些用户的性别是缺失的。 user.csv 字段顺序如下（sex性别：1 为男性，0为女性，空为缺失）：

userId, sex

 具体任务说明如下：

1. 请您抽取出user.csv 中性别完整和性别缺失两类用户，按下面的csv 格式保存, 用户性别完整的文件命名为sex1.csv, 用户性别缺失的文件命名为sex2.csv。(50分)

userId, sex

1. 选择算法，建立模型，输出建模word文档,命名为 sex\_modle.doc。(50分)
2. 编写代码，对1）中的sex1.csv文件进行训练，得到模型。为了符合开放性和创造性，解题思路可以不遵循固定的算法，不一定要输出模型，因为有些算法（比如KNN）没有模型输出。(80分)
3. 补充缺失的性别，利用训练好的模型或算法对1) 中 sex2.csv 文件的性别字段补充，补充完整后输出到 sex3.csv 文件，格式和user.csv 格式一致。(20分)