2022年重庆市大学生“大数据+”新文科应用创新大赛规程

2022.6

目录

一. 赛项属性

（一）赛项名称

2022 年重庆市大学生“大数据+”新文科应用创新大赛

（二）竞赛组织机构

主办单位：重庆市教育委员会

承办单位：西南政法大学

协办单位：重庆市统计学会

技术支持单位：悦岚（上海）数据服务有限公司

重庆坊德森科技发展有限公司

（三）组织委员会

主任委员：

温　涛　　重庆市教育委员会副主任

副主任委员：

岳彩申　　西南政法大学副校长

蒋云芳　　重庆市教育委员会高教处处长

冉渝南　　重庆市统计学会秘书长

委　员：

王怀勇　　西南政法大学教务处处长

陈　刚　　西南政法大学经济学院院长

陈　亮　　西南政法大学人工智能法学院院长

邵兵家　重庆大学经济与工商管理学院教授

何建洪　重庆邮电大学经济管理学院副院长

陈天培　重庆文理学院经济管理学院院长

颜帮全　重庆三峡学院工商管理学院院长

代　彬　　四川外国语大学经济管理学院院长

游　静　　重庆科技学院工商管理学院副院长

黄钟仪　　重庆工商大学管理科学与工程学院院长

简玉刚　　重庆工程学院管理学院院长

张　平　　重庆移通学院淬炼·国际商学院院长

张熙悦　　重庆第二师范学院经济与工商管理学院副院长

黄志平　　重庆电子工程职业学院财经管理学院院长

陈　锋　　重庆信息技术职业学院经贸学院副院长

黄菊英　　重庆城市管理职业学院数智财经学院副院长

何春华　　重庆城市职业学院商学院院长

杨　旭　　重庆财经职业学院大数据学院商务数据分析应用专业主任

王　芳　　重庆商务职业学院经贸学院副院长

高　翔　　重庆旅游职业学院财经与旅游商贸系主任

（四）专家委员会

主任委员：

余劲松　　西南政法大学经济学院党委书记

副主任委员：

哈宁武　　重庆市教育委员会高教处

薛　健　　重庆市统计学会常务理事

刘苓玲　　西南政法大学经济学院副院长

委　员：

罗先文　　西南大学商贸学院副院长

李　豪　　重庆交通大学经管学院副院长

代　应　　重庆理工大学管理学院院长

刘　进　　重庆邮电大学经管学院实验中心主任

熊　膺　　重庆师范大学经济与管理学院学院实验中心主任

罗文宝　　长江师范学院管理学院副院长

邓丽纯　　重庆城市科技学院经济管理学院副院长

李安平　　重庆对外经贸学院管理学院院长

李　亮　　重庆航天职业技术学院财经商贸学院副院长

陈现军　　重庆三峡职业学院经济管理学院副院长

罗达丽　　重庆工程职业技术学院财经与旅游学院副院长

邱　云　　重庆城市管理职业学院商学院副院长

吴　敏　　重庆青年职业技术学院经济管理学院副院长

孟　华　　重庆化工职业学院财经学院副院长

张立明　　重庆安全技术职业学院工商管理系主任

（五）仲裁委员会

主任委员：

杨　化　　西南政法大学经济学院纪委书记

委　员：

韩振国　　西南政法大学经济学院教授

黄菊英　　重庆城市管理职业学院财经学院副院长

（六）秘书处

秘书处办公室设在西南政法大学经济学院

秘书长：

李永奎　西南政法大学数字经济系主任

副秘书长：

路　瑶　西南政法大学经济学院经济系主任

张俊仪　西南政法大学经济学院办公室主任

陈玮琪　西南政法大学经济学院院团委书记

秘书：

廖雪竹　西南政法大学经济学院教务秘书

卜偲琦　西南政法大学经济学院教师

（七）赛项归属产业类型

本赛项属于交叉产业的技能竞赛，贯穿第一第二第三产业中的金融、营销、物流、财务、管理、人力资源等智能部门中和大数据应用紧密关联的复合型岗位。

（八）赛项归属专业大类/类　（更新调整）

本科	02 经济学	0201 经济学类	020101	经济学
			020102	经济统计学
			020103T	国民经济管理
			020104T	资源与环境经济学
			020105T	商务经济学
			020106T	能源经济
			020107T	劳动经济学
			020108T	经济工程
			020109T	数字经济
		0202 财政学类	020201K	财政学
			020202	税收学
		0203 金融学类	020301K	金融学
			020302	金融工程
			020303	保险学
			020304	投资学
			020305T	金融数学

			020306T	信用管理
			020307T	经济与金融
			020308T	精算学
			020309T	互联网金融
			020310T	金融科技
		0204 经济与贸易类	020401	国际经济与贸易
			020402	贸易经济
	03 法学	0301 法学类	030101K	法学
			030102T	知识产权
			030103T	监狱学
			030104T	信用风险管理与法律防控
			030105T	国际经贸规则
			030106TK	司法警察学
			030107TK	社区矫正
	05 文学	0503 新闻传播学类	050301	新闻学
			050302	广播电视学
			050303	广告学
			050304	传播学
			050305	编辑出版学
			050306T	网络与新媒体
			050307T	数字出版
			050308T	时尚传播
			050309T	国际新闻与传播
	12 管理学	1201 管理科学与工程类	120101	管理科学
			120102	信息管理与信息系统
			120103	工程管理
			120104	房地产开发与管理
			120105	工程造价
			120108T	大数据管理与应用
			120110T	计算金融

		1202 工商管理类	120201K	工商管理
			120202	市场营销
			120203K	会计学
			120204	财务管理
			120205	国际商务
			120206	人力资源管理
			120207	审计学
			120208	资产评估
			120209	物业管理
			120210	文化产业管理
			120215T	零售业管理
		1204 公共管理类	120405	城市管理
			120407T	交通管理
		1206 物流管理与工程类	120601	物流管理
			120603T	采购管理
			120604T	供应链管理
		1208 电子商务类	120801	电子商务
			120803T	跨境电子商务
		1209 旅游管理类	120901K	旅游管理
			120902	酒店管理
			120903	会展经济与管理
高职	53 财经商贸大类	5301 财政税务类	530101	财税大数据应用
			530102	资产评估与管理
		5302 金融类	530201	金融服务与管理
			530202	金融科技应用
			530203	保险实务
			530204	信用管理
			530205	财富管理
			530206	证券实务
			530207	国际金融
		5303 财务会计类	530301	大数据与财务管理

			530302	大数据与会计
			530303	大数据与审计
			530304	会计信息管理
		5304 统计类	530401	统计与大数据分析
			530402	统计与会计核算
			530403	市场调查与统计分析
		5305 经济贸易类	530501	国际经济与贸易
			530502	国际商务
		5306 工商管理类	530601	工商企业管理
			530602	连锁经营与管理
			530603	商务管理
			530604	中小企业创业与经营
			530605	市场营销
		5307 电子商务类	530701	电子商务
			530702	跨境电子商务
			530703	移动商务
			530704	网络营销与直播电商
			530705	农村电子商务
			530706	商务数据分析与应用
		5308 物流类	530801	物流工程技术
			530802	现代物流管理
			530803	航空物流管理
			530804	铁路物流管理
			530805	冷链物流技术与管理
			530806	港口物流管理
			530807	工程物流管理
			530808	采购与供应管理
			530809	智能物流技术
			530810	供应链运营
	54 旅游大类	5401 旅游类	540101	旅游管理
			540103	旅行社经营与管理

			540104	定制旅行管理与服务
			540105	研学旅行管理与服务
			540106	酒店管理与数字化运营
			540107	民宿管理与运营
			540110	智慧景区开发与管理
			540111	智慧旅游技术应用
	59 公共管理与服务大类	5901 公共事业类	590105	公共关系
		5902 公共管理类	590202	人力资源管理
			590203	劳动与社会保障
			590204	网络舆情监测

二. 赛项目的

（一）响应国家大数据战略为社会培养输送多层次“大数据+”复合型人才

本次大赛主要聚焦在文科专业在校师生的大数据思维和大数据能力的训练和培养。参加大赛的教师和学生通过全面准备、系统辅导、严格训练、积极参赛，可以在财务管理、市场营销、电子商务、物流、金融、经济学、会计、工商管理、人力资源、旅游、跨境电商、行政管理等多个文科专业掀起学习和使用的“大数据+”热潮，为社会快速培养一批高素质的复合型大数据人才，大大缓解国内企业数字化转型中对应用型和业务型大数据人才的供需矛盾。

（二）落实国务院和教育部有关高等教育的改革的最新政策

2020 年 11 月，教育部发布了《新文科建设宣言》对新文科建设作出了全面部署，新文科建设进入全面启动的新阶段。新文科建设是教育部等有关部门贯彻习近平总书记有关指示要求、顺应新时代发展趋势、推动我国高等教育内涵式发展的重要举措，新文科建设就是推动文科教育创新发展。

本次大赛覆盖的专业范围涵盖大部分的文科专业，大赛内容的设置经过充分的调研，完全贴合社会对多层次大数据人才的需求，保证竞赛中取得优异成绩的

学生，在就业中能够有较显著的竞争优势。进一步推动高校进行人才培养方案的修订和课程体系的改革，促进产教融合、校企合作和产业发展。

三. 竞赛内容

（一） 赛项名称

2022 年重庆市大学生“大数据+”新文科应用创新大赛（"Big Aata +" New Liberal Arts Application Innovation Competition of ChongQing University Students in 2022）

（二） 赛道划分

本赛项分 10 个赛道，参赛队伍可任选其中的一个赛道参加：

（1） 财务大数据

（2） 金融大数据

（3） 经济贸易大数据

（4） 工商管理大数据

（5） 电子商务大数据

（6） 物流管理大数据

（7） 旅游管理大数据

（8） 公共管理大数据

（9） 新闻传播大数据

（10） 法律大数据

（三） 竞赛内容

竞赛的内容基于行业完整的大数据应用六大环节：（1）提出业务问题；（2）框定业务数据；（3）探索业务问题；（4）开发数据流程；（5）编写业务方案；（6）解决业务问题。本赛项主要考察参赛学生的综合数据思维和数据能力,如下表：

编号	主要模块	竞赛内容
1	框定业务数据	业务问题和哪些业务数据有关联？来自哪些数据源？这些数据有哪些基本画像特征？

2	探索业务问题	通过对框定的业务数据进行探索式分析，进一步细化业务问题，把业务问题转化为技术问题
3	大数据竞赛平台基本操作	平台的基本功能 平台的大数据分析功能
4	数据思维之数据源	大数据应用场景中，各种数据源及其读写方法，包括：关系数据库、云数据库、课程数据库、云文件、课程文件、HDFS、我的数据等
5	数据思维之数据加工	大数据应用场景中，常见的数据加工方法：文本函数、数值函数、日期函数、逻辑函数、根据业务条件进行数据筛选、数据抽样、行数据缺值处理、各种数据关联加工、数据聚合加工、数据标签化等
6	数据思维之分析与挖掘	描述性统计分析、线性回归、聚类、朴素贝叶斯、关联规则等
7	数据思维之可视化	大数据应用场景中，各种数据可视化设计方法：柱形图、热力点图、面积图、文本图、地图、漏斗图、仪表盘、气泡图、雷达图等
8	综合商业数据应用案例	针对商业场景中的常见大数据应用场景，通过商业案例的方式考察学生利用数据思维和数据工具解决具体商业数据问题的能力。案例涉及到现在典型企业的财务部门、营销部门、电商部门、物流、资本金融、企业管理、人力资源等方面的业务需求。
9	大数据应用商业报告	结合综合商业案例的分析结果，从不同的业务维度出发，把数据分析的结果形成商业行动方案，指导

		各相关业务部门联动，推动企业整体目标的达成

四. 竞赛方式

（一）参赛对象

各普通高校经济学类、财政学类、金融学类、经济与贸易类、法学类、新闻传播学类、管理科学与工程类、工商管理类、公共管理类、物流管理与工程类、电子商务类和旅游管理类专业在校大学生（本科和高职高专）符合报名条件的均可报名参赛。

（二）组织方式

1.学生组以团队为单位报名，每队 3 人，其中设队长 1 人，参赛人员不得同时加入多支参赛队。学生组参赛队可配备 1～2 名教师担任指导教师，并指定 1 名指导教师作为领队。领队负责赛前辅导和参赛的组织工作。

2.参赛组队可跨年级、专业，但不得跨校，每所学校不超过 6 支队伍。

3.竞赛分本科组、高职组两个报名通道，分别参赛，各院校按对应通道组队报名参赛。

4.竞赛分初赛和决赛。初赛由各参赛学校自行组织，技术支持单位可提供技术支持，协助学校完成初赛选拔，决赛由大赛组委会统一组织。

（三）竞赛流程

大赛分为六个阶段比赛，每个赛队应该在指导教师的指导下，科学合理分工与合作，合作完成既定工作任务，彰显高等院校文科专业学生的职业能力，体现岗位通用技能、团队合作精神及职业道德等综合素质。

比赛阶段	时间	比赛方式	比赛内容	提交结果	成绩评定
互联网在线训练	2022.6.30 - 2022.9.30	在线	无	无	无
报名	2022.6.30 - 2022.9.15	在线	无	无	无
赛项启动会	2022.9.17	在线	无	无	无
初赛	2022.9.22 - 2022.10.31	在线	2022.9.17 公布初赛题目（综合性题目）	大数据工作流	由比赛系统自动评判成绩
决赛	2022.11.12	在线	2022.10.30 公布决赛题目（各赛道专业性题目）	大数据工作流+文档和报告	由比赛系统自动计算成绩
颁奖典礼	2022.11.26	在线	总结研讨会		

赛项采用统一规格的硬件、软件和辅助工具，确保竞赛平台统一、结果公平。

五.评分标准制定原则、评分方法、评分细则

按照竞赛的相关要求，结合赛项自身特点，编制赛项评分方法和评分细则。

（一）评分标准制定原则

赛项评分采用结果评分方法，始终贯彻落实大赛一贯坚持的公开、公平和公正原则。结果评分：依据赛项评价标准，对参赛选手提交的竞赛成果进行评分。赛项最终按总评分得分高低，确定奖项归属。

赛项合作企业不直接或者间接地参与赛项评分。

参与大赛赛项成绩管理的组织机构包括：裁判组、监督组和仲裁组，受赛项组委会统一领导。

1．裁判组

裁判组实行“裁判长负责制”，设裁判长 1 名，全面负责赛项的裁判与管理工作。

裁判员根据竞赛工作需要分为检录裁判和评分裁判。

检录裁判负责对参赛队伍（选手）进行姓名登记、身份核对等工作；

评分裁判负责对参赛队伍（选手）的竞赛成果按赛项评分标准进行评定。

2．监督组

监督组负责对裁判组的工作进行全程监督，并对竞赛成绩抽检复核。

3．仲裁组

仲裁组负责接受由参赛队领队提出的对裁判结果的申诉，组织复议并及时反馈复议结果。

（二）评分方法

成绩评定是根据竞赛考核目标、内容对参赛队或选手在竞赛过程中的表现和最终成果做出评价。本赛项的评分方法为结果评分。结果评分是对参赛选手提交的竞赛成果卷，依据赛项评价标准进行评价评分。所有的评分表、成绩汇总表备案以供核查，最终的成绩由裁判长进行审核确认并上报大赛组委会。

本赛项要求参赛团队根据比赛题目所提出的业务问题，从数据的角度进行全方面分析、基于业务问题分析的结果框定业务数据、基于框定的业务数据设计开

发大数据分析流程、基于大数据分析流程的输出结果撰写业务分析应用报告来指导业务部门进行有效地经营和决策。

业务分析和应用报告的主要内容包含六大模块，各模块内容要求及评价标准如下：

报告模块	评价内容	格式	权重
1 分析业务问题	1. 对业务问题的理解、解决问题的商业价值 2. 业务问题的细化 3. 从数据的角度解决业务问题的总体思路	图文	5%
2 框定业务数据	1. 业务问题和哪些业务数据有关联？来自哪些数据源？ 2. 从数据的角度解决业务问题的总体思路 3. 这些数据有哪些基本画像特征？	图文	10%
3 探索业务问题	1. 业务问题转化为什么样的技术（数据分析）问题？ 2. 发现了哪些数据质量问题？ 3. 数据流程的基本设计思路	图文	20%
4 开发数据流程	1. 数据工作流的详细设计说明书，包括大数据全链路六大环节（数据源、数据湖、数据汇集、数据加工、分析挖掘、可视化）。 2. 根据设计说明书，所开发的数据工作流	图文+大数据工作流	30%
5 编写分析报告	1. 可视化结果的导出 2. 可视化结果的业务解读	图文	25%
6 解决业务问题	把分析结果应用到业务流程中的实施方案	图文	10%

六. 奖项设置

（一） 各参赛队按照决赛成绩进行排名，且大赛按参赛队伍数量的比例颁发相关荣誉证书：

一等奖：各竞赛组参赛队伍数量的 10%，参赛选手及指导教师均可获得由大赛主办单位颁发的荣誉证书；

二等奖：各竞赛组参赛队伍数量的 20%，参赛选手及指导教师均可获得由大赛主办单位颁发的荣誉证书；

三等奖：各竞赛组参赛队伍数量的 30%，参赛选手及指导教师均可获得由大赛主办单位颁发的荣誉证书。

七. 技术规范

竞赛内容涉及技术规范的全部信息如下 ：

（一）行业标准

标准	内容
GB/T 35274-2017	信息安全技术　大数据服务安全能力要求
GB/T 35295-2017	信息技术　大数据　术语
GB/T 35589-2017	信息技术　大数据　技术参考模型

（二）软件开发工程过程标准

标准	内容
GB/Z 31102-2014	软件工程　软件工程知识体系指南
GB/T 30999-2014	系统和软件工程　生存周期管理　过程描述指南
GB/T 18726-2011	现代设计工程集成技术的软件接口规范

（三）文档标准

标准	内容
GB/T 32424-2015	系统与软件工程　用户文档的设计者和开发者要求
GB/T 8567-2006	计算机软件文档编制规范

八.建议使用的竞赛器材、技术平台要求

（一）技术平台

图形式大数据实验平台支持图形化（鼠标拖拽）的方式进行教学/科研的大数据实验。

（1）　在不需要编程序的基础上，用户可以采用鼠标拖拽设计开发 Apache Spark 大数据批式（非实时）处理作业（数据工作流）和大数据流式（实时）处理作业（流式数据工作流），可以支持大数据全链路的功能，包括数据源定义、数据汇集、数据加工、分析挖掘、数据可视化。

（2）　平台支持至少 13 种数据源功能包括：关系数据库、MongoDB、HDFS、FTP、文件夹、Hbase、系统 Hive、Cassandra、流、云文件、云数据库、课程文件、课程数据库。

（3）　平台支持至少 6 种数据转换功能：包括通用转换、流转换、自定义、分组标签、抽样、行转列。。

（4）　支持图形化调用主流统计和数据挖掘算法，包括回归、支持向量机、朴素贝叶斯、关联规则、主成份分析、PLS、层次聚类、逐步回归、随机森林、

Lasso 回归、神经网络、决策树、标准差，峰度，百分位数、移动平均、相关系数、单因素方差分析 。

（5） 平台支持 6 种数据落地功能，包括 ES 落地、HDFS 落地、Impala 落地、Hive 落地、Cassandra 落地、云文件落地。

（6） 提供调试运行功能，包括设置断点、设置目的地、校验元数据、设置查看器、设置落地的功能。启动调试任务后，可以推送相匹配的调试日志。调试日志以三类图标显示：错误信息红色图标、警告信息黄色图标、一般信息黑色图标。支持通过日志信息快速定位到关联的工作流节点的功能。

（7） 大数据平台提供至少 89 种图形化封装好的数据加工转换器组件(函数)，可以分别针对文本、数值、集合、日期数据类型的数据进行加工处理。可以通过图形式使用这些函数。支持通过鼠标右键创建数据转换器之间的输入输出连线。

（8） 支持指标 KPI 卡、迷你图、热力区域图、分区柱状图、堆积柱状图、多系列柱状图、对比柱状图、瀑布图、分区折线图、多系列折线图、折线雷达图、范围面积图、组合图、散点图、聚合气泡图、饼图、多层饼图、玫瑰图、矩形树图、词云、漏斗图等丰富的图表分析组件。提供现场演示

（9） 支持区域地图、点地图、流向地图、热力地图，可进行省、市、县级别的地图数据分析，同时支持用户直接在浏览器前端进行地图自定义设计，包括地理地图、自定义图片地图、自定义 GIS 地图背景、地理位置和经纬度自定义匹配等功能，帮助用户进行快速自定义地图编辑设计。

（10） 支持设置钻取目录，实现预览分析时对数据进行维度转换操作，满足用户从不同的视角进行数据分析查看的分析需求。

1.大数据竞赛平台系统：赛项拟采用的平台需满足以下功能要求。

2.参考硬件设备，各参赛队自备笔记本电脑，配置要求如下：

设备类型	数量	规格
客户端	每支参赛队 1 台	性能相当于 2.0GHZ 处理器,8G 以上内存，显示器要求 1920*1080 分辨率

3. 参考软件环境

设备类型	软件类别	软件名称、版本号
客户端	PC 操作系统	Windows 10
	大数据平台客户端	DEEP Client1.6
	浏览器	Chrome、火狐浏览器
	文档编辑器	Microsoft Word 2007 及以上

九.安全保障

（一）组队责任

1. 各学校代表队组成后，须制定相关管理制度，并对所有选手、指导教师进行安全教育。

2. 各参赛队伍须加强对参与竞赛人员的安全管理，实现与赛场安全管理的对接。

（二）应急处理

竞赛期间发生意外事故，发现者应第一时间报告赛项组委会，同时采取措施避免事态扩大。赛项组委会应立即启动预案予以解决并报告赛区组委会。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。事后，赛区组委会应向大赛组委会报告详细情况。

（三）处罚措施

1. 因参赛队伍原因造成重大安全事故的，取消其获奖资格。

2. 参赛队伍有发生重大安全事故隐患，经赛场工作人员提示、警告无效的，可取消其继续竞赛的资格。

3. 赛事工作人员违规的，按照相应的制度追究责任。情节恶劣并造成重大安全事故的，由司法机关追究相应法律责任。

4.认定为作弊的，则取消成绩，并报送所在院校。

（四） 赛项监督与仲裁管理

按照大赛的规定和组委会要求，完成赛项监督与仲裁管理工作。

1. 赛项监督

（1）监督组在大赛组委会领导下，对大赛筹备与组织工作实施全程监督。监督工作实行组长负责制。

（2）监督组的监督内容包括赛项竞赛时间控制赛题内容发布、竞赛纪律、成绩评判及成绩复核与发布、申诉仲裁等。

（3）监督组不参与具体赛事组织活动及裁判工作。

（4）监督组在工作期间应严格履行监督工作职责。

（5）对竞赛过程中违规现象，应及时向赛项组委会提出改正建议，同时留取监督过程资料。赛事结束后，认真填写《监督工作手册》并直接递交大赛组委会办公室存档

2．申诉与仲裁

（1）各参赛队对不符合大赛和赛项规程规定的仪器、设备、工装、材料、物件、计算机软硬件、竞赛使用工具、用品，竞赛执裁、赛场管理、竞赛成绩，以及工作人员的不规范行为等，可向赛项仲裁组提出申诉。申诉主体为参赛队领队。

（2）仲裁人员的姓名、联系方式应该在竞赛期间向参赛队和工作人员公示，确保信息畅通并同时接受大众监督。

（3）申诉启动时，参赛队向赛项仲裁工作组递交领队亲笔签字同意的书面报告。书面报告应对申诉事件的现象、发生时间、涉及人员、申诉依据等进行充分、实事求是的叙述。非书面申诉不予受理。

（4）提出申诉的时间应在竞赛结束后（选手赛场竞赛内容全部完成）2 小时内。超过时效不予受理。

（5）赛项仲裁工作组在接到申诉报告后的 2 小时内组织复议，并及时将复议结果以书面形式告知申诉方。申诉方对复议结果仍有异议，可由省（市）领队向赛区仲裁委员会提出申诉。赛区仲裁委员会的仲裁结果为最终结果。

（6）申诉方不得以任何理由拒绝接收仲裁结果，不得以任何理由采取过激行为扰乱赛场秩序。仲裁结果由申诉人签收，不能代收，如在约定时间和地点申诉人离开，视为自行放弃申诉。

（7）申诉方可随时提出放弃申诉。

具体组织分工如下：

1．设组委会主任（总指挥）一名 、副主任（副总指挥）二名，负责赛项若干事宜的总体协调。

2．设大赛秘书处：秘书长一名，组员若干，负责支持组委会主任、副主任决策的落实与监督。

3．设立仲裁组：组长一名、组员若干，负责赛项的仲裁工作。

4．设立裁判组：裁判长一名、裁判若干，负责赛项的裁判工作。

5．宣传组：组长一名，组员若干，负责赛项宣传等联系工作。

（五）赛项总结及教学研讨

以赛项总结会、研讨会等形式，传播大赛的成功经验，扩大大赛成果的影响。组织行业专家、一线教师，结合大赛题目和选手作品，共同探索竞赛目标与人才培养目标、竞赛组织与教学模式改革、实训考核与教学考核方式、职业竞赛与职业素养养成的结合方法，在兼顾知识、技能、素质发展和项目过程系统化的原则下，探索技能竞赛项目和评价标准，与专业课程项目化教学过程的有效结合，实现赛项资源向专业教学资源的转化，促进本专业教学改革。

（六）师资培训

结合大赛竞赛内容和竞赛方式，以及行业企业技能要求、教育教学需求，邀请行业技能考核专家、高等院校教学能手、企业技术专家作为培训讲师。

十.裁判人员建议

按照大赛组委会的相关要求，裁判人员的选拔标准为：

角色	专业技术方向	知识能力要求	执裁、教学、工作经历	专业技术职称（职业资格等级）	人数
评分裁判	计算机/软件工程/”大数据+”新文科专业	具备计算机专业或”大数据+”新文科专业教学经验	具有省级以上大赛执裁经验	副高及以上	5
裁判总人数	5				

十一. 竞赛试题（样卷）

“大数据+”新文科创新应用大赛样卷

考题 1（数据源与数据汇集，15 分）

某公司的销售部门记录有 2020 年的不同区域销售数据，数据在课程文件 avocado.csv，具体字段如下：

字段名	数据类型	含义
id	Int	订单 ID
total_volume	Double	每单销量
region	nvarchar	区域

请完成下列操作：

（1）从课程文件中抽取 csv 文件 avocado.csv；

（2）在（1）步的基础上分组计算不同区域的平均销量；

（3）把（2）步的计算结果落地到云数据

工作流设计需要按照以下要求命名，规定了落地表名的节点都需要落地到云数库。

1 数据工作流命名为“正式考题 1-数据源与数据湖”

2 数据源为课程文件，课程选择“‘大数据+”新文科创新应用大赛（正式题目）

3 抽取文件 avocado.csv，汇集节点命名为“抽取数据”

3 转换节点，命名为“分组求平均”，落地表名“t1_avocado_avg”，这个节点需要计算出各个地区的平均销量，形成 2 列： “region”（地区），“avg_volume”（平均销量）

考题 2（数据加工，20 分）

课程文件中的文件 shopping_records.xlsx 是两种商品在不同地区、不同店铺的销售价格、折扣、成本的数据，其中的字段 Product_name 有两个值，字段说明如下：

字段名	数据类型	含义
product_name	nvarchar	商品名称，只有两种商品：P0675、P1672
price	Double	每单销售价格
cut_off	Double	每单折扣比例
cost	Double	每单成本

请完成下列操作：

（1）设计数据空值处理逻辑，在 product_name 空值的情况下，全部补成 P1672

（2）筛选出 product_name 是 P1672 的每单销售价格、每单折扣比例、每单成本；

（3） 在（2）步的基础上，计算 P1672 的每单的毛利润；

（4）在 （3）步的基础上，计算 P1672 的平均毛利润。

工作流设计需要按照以下要求命名，规定了落地表名的节点都需要落地到云数据库

1 数据工作流命名为“正式考题 2-数据加工”

2 数据源为课程文件，课程选择“ “大数据+”新文科综合技能大赛（正式题目）”

3 抽取文件 shopping_records.xlsx，节点命名为“抽取数据”

4 转换节点，命名为“空值处理”，列的名字和类型和上一步一致。实现空值处理逻辑。

5 转换节点，命名为“筛选出 P1672”，落地表名“t2_filter_p1672”，这个节点需要筛选出产品名称为” P1672”的数据，形成 4 列： “product_name”（产品名称），“price”（价格），“cut_off”（折扣），“cost”（成本）

6 转换节点，命名为“计算 P1672 的毛利”，落地表名“t2_gross_profit”，这个节点需要计算出 P1672 的毛利，形成 1 列： “gross_profit”（毛利）

7 转换节点，命名为“计算 P1672 的平均毛利”，落地表名“t2_avg_gross_profit”，这个节点需要筛选出产品名称为 P1672 的数据，形成 1 列： “avg_gross_profit”（平均毛利）

考题 3（可视化，15 分）

创建的仪表板命名为“【账号】_正式仪表板”，如 15821234541 的用户仪表板命名为“15821234541_正式仪表板”

可视化作为大数据信息展示的重要环节，可以直观清楚地展示数据中所蕴含的信息。本考题所用数据为亚特兰大各政府部门 2017 年全年支付明细数据集，在课程文件中的 vendor_payments.csv 文件中，各字段如下：

字段名	数据类型	含义
payment_date	Nvarchar	日期
payment_amount	Double	全天付款总额（美元）

请完成下面的操作：

（1）抽取课程文件中的 vendor_payments.csv 文件；

（2）在（1）步的基础上，把表的两列落地到云数据库；

（3）在（2）步的基础上，使用三种不同的可视化形式展示付款总额（payment_amount）随付款日期（payment_date）的变化情况。

考题 4（数据挖掘，20 分）

吉利汽车公司希望通过在美国设立生产部门并在当地生产汽车，从而在美国和欧洲同行中竞争来进入美国市场。

本考题的数据集有两个，car_price_train.csv 和 car_price_predict.csv。其中 car_price_train.csv 是训练数据，car_price_predict.csv 是预测数据。训练数据 car_price_train.csv 只是比预测数据 car_price_predict.csv 多了标签列，其它列完全相同。

训练数据 car_price_train.csv 的信息如下表所示：

字段名	数据类型	含义
car_id	Int	编号
symboling	Int	保险风险评级（值为+3 表示汽车有风险，-3 表示相当安全）
wheelbase	Double	汽车轴距
carlength	Double	车辆长度
carwidth	Double	轿厢宽度
carheight	Double	轿厢高度
curbweight	Int	汽车的重量
enginesize	Int	车辆尺寸
boreratio	Double	车辆钻孔率
stroke	Double	发动机容积
compressionratio	Double	汽车压缩比
horsepower	Int	马力
peakrpm	Int	车辆转速峰值
citympg	Int	每 100 公里城市油耗
highwaympg	Int	每 100 公里高速油耗
price	Double	价格。标签列

请完成下列操作：

（一）利用训练数据 car_price_train.csv 训练一个线性回归模型，该模型能够对汽车的价格进行预测。工作流设计需要按照以下要求命名，规定了落地表名的节点都需要落地到云数据库

1 数据工作流命名为“正式考题 4-数据挖掘-训练”

2 数据源为课程文件，课程选择“‘‘大数据+”新文科创新应用大赛（正式题

目）”

3 抽取文件 car_price_train.csv，节点命名为“抽取数据”

4　转换节点，命名为“训练数据特征组装”，落地表名“t4_car_price_train_csn”，这个节点需要将训练数据组装成 csn 格式的，形成 1 列：　“train_csn”（数据组装）

5　线性回归节点，这个节点需要完成模型训练

（二）利用（1）步训练的模型在预测数据集 car_price_predict.csv 上预测每辆车的价格。工作流设计需要按照以下要求命名，规定了落地表名的节点都需要落地到云数据库

1 数据工作流命名为“正式考题 4-数据挖掘-预测”

2 数据源为课程文件，课程选择“‘大数据+”新文科综合技能大赛（正式题目）”

3 抽取文件 car_price_predict.csv，节点命名为“抽取数据”

4　转换节点，命名为“预测数据特征组装”，落地表名“t4_car_price_predict_csn”，这个节点需要将预测数据组装成 csn 格式的，形成 2 列：　“car_id”（汽车 id），　“predict_csn”（数据组装）

5　转换节点，命名为“预测”，落地表名“t4_car_price_predict”，这个节点需要预测出车辆的价格，形成两列：　“car_id”（汽车 id），　“predict_price”（预测价格）

考题 5（案例操作，30 分）

某跨国公司的人力资源部汇集了大量员工的信息数据，包括企业因素如（部门 sales）、员工行为相关信息（项目数、平均每月工作时长、工作年限、是否升值，工资水平等），以及工作相关因素（员工满意程度、最新绩效评估、是否出现工作事故）。这些因素有利于分析员工流失原因以及预测在岗员工是否可能流失，可以采取措施挽留具有流失倾向的员工。

本考题的数据集是：dimission.csv，其元数据如下表所示：

字段名	数据类型	含义
id	Int	员工编号
satisfaction_level	Int	员工对公司满意度
last_evaluation	Int	最新评价
number_project	Int	项目数
average_montly_hours	Int	平均每月工作时长
time_spend_company	Int	工作年限

work_accident	Int	是否出现工作事故
dimission	Int	标签列，是否离职（0 表示未离职，1 表示离职）
promotion_last_5years	Int	过去 5 年是否升职
sales	Nvarchar	岗位，10 个值，分别是 (IT 、 management 、 support 、 product_mng、RandD、technical、sales、hr、accounting、marketing)
salary	Nvarchar	薪资水平，三个值，分别是(low、medium、high）

请完成下列操作：

工作流设计需要按照以下要求命名，规定了落地表名的节点都需要落地到云数据库

1 数据工作流命名为“正式考题 5-综合案例”。

2 数据源为课程文件，课程选择“ “大数据+”新文科创新应用大赛（正式题目）”。

3 抽取文件 dimission.csv，抽取后的结果节点命名为“抽取数据”

4 对原始业务数据的文本型数据进行加工，通过 Choice 转换器将文本类型的数据转换成数值型数据。转换后的结果节点命名为“字符串转数值类型”，落地表名“t5_string_to_num”，形成 10 列：

列名	数据类型
satisfaction_level	Double
last_evaluation	Double
number_project	Int
average_montly_hours	Int
time_spend_company	Int
work_accident	Int
dimission	Int
promotion_last_5years	Int
sales	Int
salary	Int

5 将数据的 average_montly_hours 进行离散化，离散化后的列的名称与上一步相同。离散后的节点命名为“工作小时离散化”，落地表名

“t5_dimission_hours”。

6 讲上步结果的 70%分为训练数据，节点命名为“训练数据”，剩余的 30%分为测试数据，节点命名为“测试数据”。

7 将训练数据组装成 csn 格式的，形成 1 列： “train_csn”（数据组装）。组装后的节点命名为“训练数据特征组装”，落地表名“t5_dimission_train_csn”，

8 训练一个逻辑回归分类模型。

9 将测试数据组装成 csn 格式的，形成 2 列： “test_csn”（数据组装），“dimission”（是否离职）。组装后的节点命名为“测试数据特征组装”，落地表名“t5_dimission_test_csn”，

10 对测试数据调用第 8 步中训练的模型，节点命名为“预测结果”,落地表名命名为“t5_dimission_result”，包含两列：“dimission_predict”(预测结果)和“dimission”（原始标签）。