全国职业院校技能大赛
赛项规程

一、赛项名称

赛项编号：GZ-2021041

赛项名称：大数据技术与应用

英文名称：Big Data Technology And Application

赛项组别：高职

赛项归属：电子信息大类

二、竞赛目的

为推进实施国家大数据战略，进一步落实国务院《关于促进大数据发展行动纲要》（国发[2015]50号）以及2021年政府工作报告中“加快数字化发展，打造数字经济新形态，协同推进数字产业化和产业数字化转型，加快数字社会建设步伐，提高数字政府建设水平，营造良好数字生态，建设数字中国。”等要求，不断推进大数据专业人才培养，建立健全多层次、多类型的大数据人才培养体系。

本赛项旨在结合当前大数据行业中技术要求，通过大赛让参赛选手熟悉一个大数据项目中各个环节的实现过程。通过竞赛来检验教学水平，引领和促进职业教育教学改革，促进与世界最新水平接轨，营造崇尚技能的社会氛围。

通过大赛培养参赛选手在企业真实项目环境下进行大数据平台部署管理、数据采集与处理、数据清洗与挖掘分析、数据可视化及综

合分析的能力；同时培养选手的理解力、沟通力、抗压力、6S 规范等职业素质；激发学生自主学习能力和解决问题能力，以达到“以赛促学、以赛促教、以赛促改”目的。

赛项围绕大数据产业各个岗位的实际需求和要求进行设计，通过大赛搭建校企合作的平台，深化产教融合，推进产教融合人才培养模式，提升大数据技术与应用专业及其他相关专业毕业生能力，同时大赛将促进相关教材、资源、师资、认证、实习就业等全面建设，推动院校和企业联合培养大数据人才，加强学校教育与产业发展的有效衔接，促进职业院校信息类相关专业共同发展，为国家战略规划提供大数据领域的高素质技能型人才。

三、竞赛内容

（一）选手需具备能力

本赛项基于企业真实项目和工作模块，结合企业岗位对学生职业技能的最新需求，在规定的时间内完成指定大数据模块。其中，主要考核参赛选手在大数据平台部署管理、数据采集与处理、数据清洗与挖掘分析、数据可视化及综合分析等方面技能。此外，竞赛同时考核参赛选手工作组织和团队协作能力、沟通和人际交往能力、解决问题能力以及致力于紧跟行业发展步伐的自我学习能力。

本项目竞赛内容通过对技能实操表现来评估知识理解以及技能的熟练程度，将不再另外举行知识及理解性质的理论测试。

（二）竞赛模块

1. 竞赛时间

竞赛总时长为 8 小时。各竞赛队在规定的时间内，独立完成“竞赛内容”规定的竞赛模块。

2.竞赛内容

本竞赛结合国内行业、企业的实际业务模型；本竞赛只考核技能部分，不涉及理论。本竞赛进行的技能实操考核，涉及大数据平台及组件的部署管理、数据采集与处理、数据清洗与挖掘分析、数据可视化、综合分析。

序号	比赛模块	分数占比	考核内容
1	Hadoop 平台及组件的部署管理	15%	选手对 Hadoop 平台及组件的部署、管理、应用
2	数据采集与处理	20%	选手多维度数据采集能力，包括对关系型数据库、非关系型数据库和网络爬虫技术的应用
3	数据清洗与挖掘分析	25%	选手对 Hadoop 系统、Spark 系统、数据仓库等综合应用能力，使用 Java、Python、Scala 等开发语言，完成数据清洗、数据存储、数据转化、数据分析、数据挖掘等操作
4	数据可视化	20%	选手通过常见的数据可视化方法，使用 Python 语言将数据分析结果以图表的形式进行呈现、统计
5	综合分析	15%	选手对大数据技术与分析的综合操作和业务分析及报告撰写能力
6	职业素养	5%	团队分工明确合理、操作规范、文明竞赛

备注 1：关于最终赛题将由专家组讨论决定。其中，各模块的详细内容描述如下：

（1）Hadoop 平台及组件的部署管理

依据大数据 Hadoop 平台的技术特点能够独立下载、解压、安装 Hadoop 不同版本的介质。能够对不同版本的 Hadoop 介质进行文件参数配置，日志查看、状态查看、服务启动、组件部署、组件管理等。

参赛选手需要掌握以下并不仅限于以下技能:

- 根据需求解压一个 Hadoop 版本
- 根据需求设置一个 Hadoop 环境变量
- 根据需求配置相关 Hadoop 环境文件
- 根据需求配置相关 Hadoop 环境状态
- 根据需求管理相关 Hadoop 环境启停
- 根据需求部署 Hive 组件
- 根据需求管理 Hive 组件
- 根据需求应用 Hive 组件
- 根据需求部署 HBase 组件
- 根据需求管理 HBase 组件
- 根据需求应用 HBase 组件
- 根据需求部署 Sqoop 组件
- 根据需求管理 Sqoop 组件
- 根据需求应用 Sqoop 组件
- 根据需求部署 Kafka 组件
- 根据需求管理 Kafka 组件
- 根据需求应用 Kafka 组件
- 根据需求部署 Flume 组件

- 根据需求管理 Flume 组件
- 根据需求应用 Flume 组件
- 根据需求部署 Spark 组件
- 根据需求管理 Spark 组件
- 根据需求应用 Spark 组件
- 根据需求部署 Zookeeper 组件
- 根据需求管理 Zookeeper 组件
- 根据需求应用 Zookeeper 组件

（2）数据采集与处理模块

利用 Chrome 浏览器查看网页源码、分析网站网页结构。按照要求使用 Python 语言编写爬虫代码、爬取指定数据项；综合利用 ETL 工具对企业数据进行采集，并对采集结果数据集进行数据探索、以及必要的处理操作。

参赛选手需要掌握以下并不仅限于以下技能:

- 能够按要求对网页源码进行分析
- 能够按要求分析出网页结构
- 能够创建爬虫项目框架
- 能够按要求构建爬虫请求
- 能够按要求定义相关字段
- 能够按要求获取有效数据
- 能够将爬取到的数据保存到本地
- 能够将爬取到的数据保存到指定 Mysql

- 能够将爬取到的数据保存到 HDFS
- 能够将爬取到的数据保存到 HBase
- 能够对爬取到数据集进行数据探索
- 能够对爬取到的数据进行必要处理

（3）数据清洗与挖掘分析

利用 Java、Python、Scala 等开发语言，根据数据的缺失、分布等基本情况，完成数据清洗、数据转化工作；并根据实际业务需求，根据既有数据模型完成数据分析、数据挖掘操作。

参赛选手需要掌握以下并不仅限于以下技能:

- 根据需求配置 Java 环境变量
- 根据需求配置 Java 编译环境 IDEA
- 根据需求创建 Java 工程项目
- 根据需求配置 Java 工程项目
- 根据需求应用 Java
- 根据需求安装 Python
- 根据需求配置 Python 环境变量
- 根据需求安装 Python 编译环境 PyCharm
- 根据需求配置 Python 编译环境
- 根据需求应用 Python 第三方库
- 根据需求配置 Python 方法参数
- 根据需求安装 Scala 编译环境 IDEA
- 根据需求配置 Scala 编译环境

- 根据需求应用 Spark 第三方库
- 根据需求配置 Spark 方法参数
- 根据需求进行数据排序
- 根据需求进行数据集成
- 根据需求进行数据筛选
- 根据需求处理缺失数据
- 根据需求处理数据极值
- 根据需求进行数据平滑处理
- 根据需求进行数据分箱操作
- 根据需求进行数据变换
- 根据需求建立聚类模型
- 根据需求建立逻辑回归模型
- 根据需求建立决策树分类模型
- 根据需求建立随机森林分类模型
- 根据需求建立神经网络模型
- 根据需求建立支持向量机模型
- 根据需求建立线性回归模型
- 根据需求设置数据模型参数
- 根据需求进行分类模型评估
- 根据需求进行回归模型评估
- 根据需求进行聚类效果评估
- 根据需求进行模型参数优化

（4） 数据可视化

对数据进行统计、分析，并利用 Flask 开源引擎、Jinja2 模板引擎以及常用的数据可视化工具，如：ECharts，将数据分析结果以柱状图、饼图、条形图等图表进行展示。

参赛选手需要掌握以下并不仅限于以下技能:

- 根据需求应用 Flask 框架
- 根据需求应用 Jinja2
- 根据需求创建 ECharts 项目
- 根据需求使用 ECharts 绘制柱状图
- 根据需求使用 ECharts 绘制堆叠柱状图
- 根据需求使用 ECharts 绘制基础折线图
- 根据需求使用 ECharts 绘制折线堆叠图
- 根据需求使用 ECharts 绘制瀑布图
- 根据需求使用 ECharts 绘制折柱混合图
- 根据需求使用 ECharts 绘制玫瑰图
- 根据需求使用 ECharts 绘制数据聚合图
- 根据需求使用 ECharts 绘制气泡图
- 根据需求使用 ECharts 绘制地理坐标图
- 根据需求使用 ECharts 绘制 K 线图
- 根据需求使用 ECharts 绘制盒须图
- 根据需求使用 ECharts 绘制饼图
- 根据需求使用 ECharts 绘制条形图

- 根据需求使用 ECharts 绘制雷达图
- 根据需求使用 ECharts 绘制热力图
- 根据需求使用 ECharts 绘制关系图
- 根据需求使用 ECharts 绘制漏斗图
- 根据需求使用 ECharts 绘制仪表盘

（5）综合分析模块

依据网站分析及数据爬取、数据清洗与挖掘分析及可视化呈现，在综合理解业务数据的基础上，根据题目要求进行分析，并编写输出分析报告。

参赛选手需要掌握以下并不仅限于以下技能:

- 根据要求结合聚类算法结果进行可视化呈现，并说明聚类对业务发展的用途及经营策略影响
- 根据要求结合归一算法结果进行可视化呈现，并说明归一对业务发展的用途及经营策略影响
- 根据要求结合排序算法结果进行可视化呈现，并说明排序对业务发展的用途及经营策略影响
- 根据要求结合决策树算法结果进行可视化呈现，并说明决策树对业务发展的用途及经营策略影响
- 根据要求结合归一算法结果进行可视化呈现，并提出合理化建议
- 根据要求结合排序算法进行可视化呈现，并提出合理化建议
- 根据要求结合决策树算法进行可视化呈现，并提出合理化建议

- 根据要求结合业务背景及相关分析结论，对未来业务规划提出建议

四、竞赛方式

1.本赛项为团体赛，以院校为单位组队参赛，不得跨校组队。每支参赛队由3名选手（设队长1名）且均为在籍高职院校学生。其中，参赛选手年龄须不超过25周岁(年龄计算的截止时间以2021年5月1日为准)，其性别和年级不限。指导教师须为本校专兼职教师，每支参赛队不超过2名指导教师;

2.本赛项设单一场次，所有参赛队在现场根据给定的项目模块，在8小时内相互配合，采用小组合作的形式完成赛项模块，最后以提交的截图和文档作为最终评分依据;

3.不计参赛选手的个人成绩，统计竞赛队的总成绩进行排序。

五、竞赛流程

（一）竞赛流程图

2021年大数据技术与应用赛项的竞赛流程如图1所示。

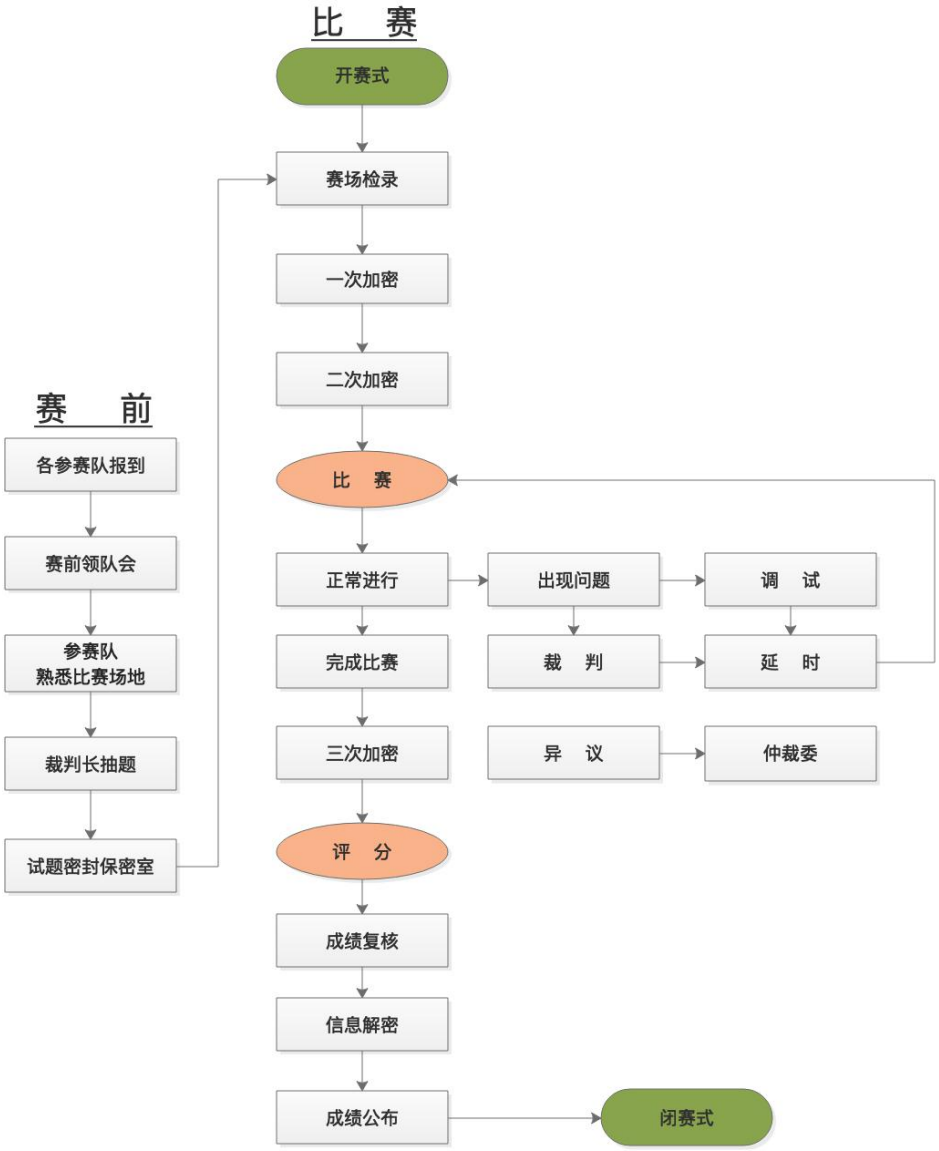

图 1　竞赛流程图

（二）竞赛时间表

日期	时间	内容
比赛前 2 日	18: 00 之前	裁判报到
	19: 00—20: 00	裁判工作会议
比赛前 1 日	12: 00 之前	各参赛队报到
	10: 00—11: 00	工作人员（含监考）培训会
	15: 30—16: 00	赛前领队会

	16:00—16:30	参赛队熟悉比赛场地
	17:00—18:00	现场裁判赛前检查，封闭赛场
比赛当日	07:00—08:00	参赛队集合前往比赛现场
	08:00—08:10	赛场检录
	08:10—08:30	一次加密：参赛队抽取参赛编号
	08:30—08:45	二次加密：参赛队抽取赛位号
	08:45—09:00	参赛队进入比赛赛位，进行赛前软、硬件检查、题目发放
	09:00—17:00	比赛
	17:00—17:20	收取各参赛队赛题及比赛结果文档
	17:00—19:00	申诉受理
	19:00—19:30	三次加密：竞赛结果等文件加密
	19:30—23:00	成绩评定与复核
	23:00—23:30	加密信息解密
	23:30—24:00	成绩汇总，报送及公布
比赛后 1 日	09:00—10:00	闭赛式

六、竞赛赛卷

（一）专家组建立赛卷库

本赛项建立竞赛赛卷库，样题由全国职业院校技能大赛执委会组织专家组完成，并基于全国职业院校技能大赛相关文件相关技术文件要求，完成竞赛赛卷库建设。制作完成的竞赛赛卷库于开赛前 1 个月，通过大赛信息发布平台进行公开。其中，竞赛样卷与竞赛规程同步发布。

（二）裁判长确定赛题

基于已经公布的竞赛赛卷库，赛前裁判长指定相关人员抽取其中 2 套赛卷。专家组将每套赛卷中 30%内容进行重新编制，最终形成 2

套正式赛题，作为竞赛用的 A 卷与 B 卷，并封存于承办院校保密室中。保密室全程监控，并安排专人把守。正式赛卷在比赛前一天，由裁判长指定的人员在监督仲裁长监督下，从正式赛题库中随机抽取竞赛试题。

比赛完成后，包括参赛选手在内的任何人，都不得将赛题带离赛场，由现场裁判对赛题进行回收。

赛卷样式具体参考样卷，见附件。

七、竞赛规则

1. 参赛队及参赛选手资格。参赛选手须为高职院校全日制在籍注册学生、本科院校中高职类全日制在籍注册学生、五年制高职四、五年级在籍注册学生。参赛选手年龄须不超过 25 周岁(年龄计算的截止时间以 2021 年 5 月 1 日为准)。凡在往届全国职业院校技能大赛中获本赛项高职组一等奖的选手，不能再报名参赛。

2. 比赛工位通过抽签决定，比赛期间参赛选手原则上不得离开比赛场地。参赛选手按规定时间到达指定地点，凭参赛相关凭据进入赛场。选手迟到 10 分钟取消比赛资格。

3. 竞赛所需的硬件、软件和辅助工具统一提供，选手不得私自携带任何移动存储、辅助工具、移动通信等设备进入赛场。

4. 参赛选手在赛前 15 分钟进入比赛工位，并由队长领取比赛信息。比赛正式开始后方可进行相关操作。如出现较严重的违规、违纪、舞弊等现象，经裁判组裁定取消比赛成绩。

5. 在比赛过程中，参赛选手如有疑问，应举手示意，现场裁判应

按要求及时予以答疑。如遇设备或软件等故障，参赛选手应举手示意，现场裁判、技术人员等应及时予以解决。确因计算机软件或硬件故障，致使操作无法继续，经裁判长确认，予以启用备用设备。

参赛选手不得因各种原因提前结束比赛。如确因不可抗因素需要离开赛场的，须向现场裁判员举手示意，经裁判员许可并完成记录后，方可离开。凡在竞赛期间内提前离开的选手，不得返回赛场。

6. 比赛时间结束，选手应全体起立，结束操作。经工作人员查收清点所有文档后方可离开赛场，离开赛场时不得带走任何资料。

7. 赛项裁判应严格遵守赛项各项规章制度，确保比赛公平、公正、公开。比赛当天 8:00 起，赛项裁判应上交所有通信设备，由赛项执委会统一保管，并安排赛项裁判在指定区域休息或工作，直至赛项成绩评定结束。

8. 比赛结束，经加密裁判对各参赛选手提交的竞赛结果进行第三次加密后，评分裁判方可入场进行成绩评判。

最终竞赛成绩经复核无误，由裁判长、监督仲裁长签字确认后，以纸质形式向全体参赛队进行公布，并在闭赛式上予以宣布。

9. 本赛项各参赛队最终成绩，由承办单位信息员在监督仲裁组监督下录入赛务管理系统。承办单位信息员对成绩数据审核后，将赛务系统中录入的成绩导出打印，经赛项裁判长审核无误后，签字。

承办单位信息员将裁判长确认的电子版赛项成绩上传赛务管理系统；同时，将裁判长签字的纸质打印成绩单报送大赛执委会。

10. 赛项结束后，专家工作组根据裁判判分情况，分析参赛选手

在比赛过程中对各知识点、技术的掌握程度，并将分析报告报备大赛执委会办公室，执委会办公室根据实际情况适时公布。

11.赛项中每个比赛环节裁判判分的原始材料和最终成绩等结果性材料，经监督仲裁组人员和裁判长签字后，装袋密封留档；并由赛项承办院校封存，委派专人妥善保管。

12.其它未尽事宜，将在赛前向各领队做详细说明。

八、竞赛环境

（一）赛场布局要求

竞赛现场设置场内竞赛区、裁判工作区、技术支持区、服务区等

1.场内竞赛区域。每个竞赛工位标有醒目的工位编号，每个工位面积在 9 ㎡左右，工位之间由隔板隔开，确保参赛队之间互不干扰。赛场要求竞赛过程全程无死角视频监控，监控录像保存 3 个月。环境标准要求保证赛场采光（大于 500 lux）、照明和通风良好；提供稳定的水、电，并提供应急的备用电源；提供足够的干粉灭火器材。

2.裁判工作区。供裁判休息及工作场地。共配有电脑 10 台；A4 激光打印机 1 台；桌椅 10 套；饮水机、纸杯、文具用品若干。

3.技术支持区。为技术支持人员的工作场地，为参赛选手竞赛提供技术支持。

4.服务区。提供医疗等服务保障，并用隔离带隔离。

（二）赛事安全要求

1.禁止选手及所有参加赛事的人员，携带任何有毒有害物品进入竞赛现场。场内竞赛区为参赛队提供统一的竞赛设备，无需选手自带任何工具及附件。

2.承办单位应设置专门的安全防卫组，负责竞赛期间健康和安全事务。主要包括检查竞赛场地、与会人员居住地、车辆交通及其周围环境的安全防卫；制定紧急应对方案；监督与会人员食品安全与卫生；分析和处理安全突发事件等工作。

3.赛场须配备相应医疗人员和急救人员，并备有相应急救设施。

4.承办方应按照疫情防范要求做好赛场各项工作，现场消防器材和消防栓合格有效，应急照明设施状态合格，赛场明显位置张贴紧急疏散图，赛场地面张贴荧光疏散指示箭头，赛场出入口专人负责，随时保证安全通道的畅通无阻。

九、技术规范

本赛项的技术规范将包括：相关专业的教育教学要求、行业、职业技术标准，以及根据高职目录修订后的大数据技术与应用相关专业人才培养标准和规范，适时地修订本赛项遵循的技术规范。

1.基础标准

标准号/规范简称	名称
GB/T 11457-2006	信息技术、软件工程术语
GB8566-88	计算机软件开发规范
GB/T 12991-2008	信息技术数据库语言 SQL 第 1 部分：框架
GB/T 21025-2007	XML 使用指南
GB/T 28821-1012	关系数据管理系统技术要求
LD/T81.1-2006	职业技能实训和鉴定设备技术规范

2.大数据技术相关标准

标准号/规范简称	名称
GB/T 38672-2020	信息技术 大数据接口基本要求

GB/T 38673-2020	信息技术 大数据大数据系统基本要求
GB/T 38676-2020	信息技术 大数据存储与处理系统功能测试要求
GB/T 38643-2020	信息技术 大数据分析系统功能测试要求
GB/T 38675-2020	信息技术 大数据计算系统通用要求
GB/T 38633-2020	信息技术 大数据系统运维和管理功能要求

3.软件开发标准

标准号/规范简称	名称
GB/T 8566 -2001	信息技术软件生存周期过程
GB/T 15853 -1995	软件支持环境
GB/T 14079 -1993	软件维护指南
GB/T 17544-1998	信息技术软件包质量要求和测试

十、技术平台

（一）竞赛设备

技术平台软硬件设备组成如下:

序号	设备名称	数量	备注
1	服务器	1	支撑大数据竞赛管理系统运行使用。内嵌基于 KVM 开发虚拟化软件，要求该软件提供虚拟机快照、一键切换大屏、存储一键清理、纵向 DRX、虚拟机回收站、在线克隆为模板、批量修改虚拟机的配置参数。作为虚拟化资源管理系统的计算资源、网络资源和存储资源的源节点。 1. CPU 模块: 2*2.3GHz 2. 内存模块: 8*32GB 3. 硬盘模块: 6*600GB SAS 10K 4. 网口: 4 端口千兆电接口网卡-360T-B2 5. 1+1 冗余电源

2	大数据竞赛平台	1	1.系统基于Linux系统部署，支持多工作节点分布式部署模式、多角色管理、赛前设定试卷自动下发、数据统计、赛后收集数据信息、现场答题、在线提交试卷功能；支持通过VNC、console、rdp多种模式访问竞赛平台；
			2.支持自动评分；支持程序代码、系统命令、运行结果、运行日志等全方位自动化评分；支持答卷脏数据清洗和异常字符规整、自动评分文件支持AES、多组评分规则导入和多套赛题适配；
			3.支持模拟大数据环境搭建、大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现和应用等贯穿大数据技术的相关知识点，提供大数据竞赛管理系统所需的虚拟服务器，所涉及开发语言包括Java、Python、Scala、HTML、Javascript等。
3	PC机	3	竞赛选手比赛使用。性能相当于i5处理器，8G以上内存，1TB以上硬盘，显示器要求1024*768以上。
4	交换机	1	1.机架式交换机 2.端口：⩾24个10/100/1000Base-TX以太网端口； 3.速度：10/100/1000Base； 4.全千兆三层交换机，支持访问控制。

备注：实际赛场需要的服务器、PC机和交换机数量取决于参赛队伍数量。

（二）软件环境

设备类型	软件类别	软件名称、版本号
服务器集群	大数据集群操作系统	CentOS 7.4
	大数据分析平台组件	Hadoop 2.6.0
		Yarn 2.6.0
		Zookeeper 3.4.5
		Hive 1.1.0

		Flume 1.6.0
		Sqoop 1.4
		kafka 1.0
		Spark 2.0
	数据库	MySQL 5.7
开发客户端	PC 操作系统	Windows 10 64 位
	浏览器	Chrome
	开发语言	Python 3.6 64bit
		Java 8
		Scala 11
	开发工具	Pycharm 2019 （Community Edition）
		IDEA 2019 （Community Edition）
	数据采集组件	Requests
		Scrapy
	数据可视化组件	ECharts 4.0
		Flask
		Jinja2
		Matplotlib
	文档编辑器	WPS 2019 及以上
	输入法	拼音输入法

注意：未来竞赛中大数据平台组件将升级为 Hadoop3.0 及以上和配套兼容组件。

十一、成绩评定

（一）评分原则

客观性试题裁判不参与评分，通过参赛选手提交的结果，由后台

竞赛系统进行自动评分，体现客观试题评分的准确性、全面性和公平性。主观性试题由多名裁判共同评分，计算各裁判打分的平均值，作为该模块的最终分数。

1.评分表样例

评分表按照选手对应题目要求实现过程及结果进行评分，具体评分样表如下。

模块	任务	主要知识与技能点	分值
模块 A: Hadoop 平台及组件的部署管理	任务一: Hadoop 伪分布部署管理	Hadoop 伪分布下的 JDK 的解压安装、JDK 环境变量配置、Hadoop 解压、Hadoop 环境变量配置、Hadoop 配置文件修改、Hadoop 启停状态	7
	任务二： Hive 组件部署管理	Hive 的解压安装、Hive 的环境变量配置、hive-site.xml 文件配置、初始化 Hive 元数据、Hive 的启停状态	4
	任务三： Kafka 组件部署管理	Kafka 的解压安装、Kafka 的环境变量配置、Kafkaserver.properties 文件修改、Kafka 启停，kafka 结果输出	4
	小计		15
模块 B: 数据采集与处理	任务一: 网页源码应字段	使用 Chrome 浏览器，查找网站异步请求的数据，并将正确内容及答案完整复制粘贴至报告中	4
	任务二: 自行创建 Scrapy 工程	使用 Python 程序自行创建 Scrapy 工程编写正确爬虫代码，以合理的程序逻辑判断相关数据包含的页数并将程序代码复制粘贴至对应报告中	2
	任务三： 在	根据爬取字段，在 MySQL 中创建 crawl 数据库，在该数据库中创建 accommodations1	4

	MySQL 中创建数据库表	表、accommodations1 表	
	任务四: 对数据库表排序	在数据库中，能正确对 accommodations1 表降序排序	4
	任务五: 对数据表填充处理	在数据库中，能正确对 accommodations1 表填充处理	3
	任务六: 对数据表删除处理	在数据库中，能正确对 accommodations2 表删除处理	3
	小计		20
模块 C: 数据清洗与挖掘分析	任务一: 数据清洗	对从企业消费平台获取到的数据进行缺失值、空值删除等编辑程序实现数据清洗，并显示数据清洗后的结果	11
	任务二: 数据挖掘分析	通过编程程序和算法，对数据清洗后的数据集进行数据挖掘分析	14
	小计		25
模块 D: 数据可视化	任务一: 柱状图呈现城市出租率	正确使用 Flask 框架，结合 Echarts 绘制柱状图	3
	任务二: 折线图呈现连锁住宿场所出租率	正确使用 Flask 框架，结合 Echarts 绘制折线图	3
	任务三: 散点地图呈现各城市住宿场所间夜数	正确使用 Flask 框架，结合 Echarts 绘制散点地图	3
	任务四: 堆叠柱状图呈现直销和分销直销和	正确使用 Flask 框架，结合 Echarts 绘制堆叠柱状图	3

	分销		
	任务五：使用sklearn库中方法构建线性回归图	正确使用Flask框架，结合Echarts绘制线性回归图	4
	任务六：多线雷达图呈现各省份住宿场所综合情况	正确使用Flask框架，结合Echarts绘制多线雷达图	4
	小计		20
模块E：综合分析	任务一：通过数据及图示分析原因	正确以各省住宿场所出租率和各城市住宿场所间夜数的折线图，对各省住宿场所的运营情况进行分析	5
	任务二：对通过图示和计算业务分析原因	正确以模块D可视化分析对某连锁酒店在不同地区的酒店出租率的统计，说明影响酒店出租率的原因	5
	任务三：对企业消费平台未来拓展合作建议和意见	正确提出企业消费平台未来拓展合作住宿场所的方向提出建议	5
	小计		15
模块F：职业素养	考察职业素养	竞赛团队分工明确合理、操作规范、文明竞赛	5
	小计		5
总分			100

2. 三次加密原则

比赛过程采取三次加密，通过抽取参赛编号、工位号和竞赛成果号，屏蔽参赛队信息，每个环节设置一名独立裁判，每个环节结束后，数据立即封存于承办校保密室保险柜内，加密裁判直接隔离，确保成绩评定公平、公正。

3. 独立评分原则

根据裁判分工，负责相同模块评分工作的不同裁判，采取随机抽签独立评分，确保成绩评定严谨、客观、准确。裁判进行随机抽签分组，杜绝主观意愿组队，各自完全独立评分，裁判员间互不干涉，比赛监督人员可随机监督。

4. 错误不传递原则

各模块分别计算得分，错误不传递，按规定比例计入选手总分。

5. 抽查复核原则

（1）为保障成绩评判的准确性，监督仲裁组对赛项总成绩排名前 30%的所有参赛队伍（选手）的成绩进行复核；对其余成绩进行抽检复核，抽检覆盖率不得低于 15%。

（2）监督仲裁组需将复检中发现的错误以书面方式及时告知裁判长，由裁判长更正成绩并签字确认。

（3）复核、抽检错误率超过 5%的，则认定为非小概率事件，裁判组需对所有成绩进行复核。

（二）评分方法

1.竞赛满分为100分。最终成绩按100分制进行排名。

2.竞赛采取三次加密。第一次加密裁判组织参赛选手第一次抽签，抽取参赛编号，替代选手参赛证等个人信息；第二次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；第三次加密裁判对各参赛队竞赛结果进行加密，替换赛位号。每个环节结束后，数据立即封存于承办校保密室保险柜内，加密裁判直接隔离，在评分结束后进行解密并统计成绩。

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层解密：竞赛结果编号到工位号解密；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，监督仲裁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. 监督仲裁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。

（三）裁判要求

序号	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

1	信息技术	信息技术大类	执裁过省级竞赛，教授过信息技术相关课程	副教授及以上	1
2	信息技术	信息技术大类	执裁过省级竞赛，教授过信息技术相关课程	副教授及以上	10
3	信息技术	大数据	执裁过省级竞赛，教授过大数据相关课程	副教授及以上	9
4	无	无	无	副教授及以上	3
裁判总人数	竞赛设置裁判 23 人，包括裁判长 1 名，裁判 17 名。其中加密裁判 3 人，现场裁判 10 人，评分裁判 9 人				

注意：承办校可根据本校场地实际情况增加现场裁判数量。

十二、奖项设定

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）。

如出现参赛队总分相同情况，按照模块分值权重顺序的得分高低排序，即总成绩相同的情况下比较模块 C 的成绩，模块 C 成绩高的排名优先，如果模块 C 成绩也相同，则按模块 D、模块 B、模块 A、模块 E 模块的成绩进行排名，以此类推完成相同成绩的排序。如果所有模块分值相同，则查看文档撰写规范、职业素养的分值进行排序。

获得一等奖的参赛队的指导教师获“优秀指导教师奖”。

十三、赛场预案

（一）应急安全预案

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。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，是否停赛由赛区执委会决定。事后，赛区执委会应向大赛执委会报告详细情况。

相关应急预案如下表所示。

突发事件	预防措施	事件发生后应对措施
参赛选手发病或受伤	在各工位张贴安全操作说明。	医务人员应采取紧急救护措施，及时进行救治，如病情或伤势严重，应及时送往最近医院进行救治。
人员发生食物中毒	比赛期间指定的住宿/餐饮场地符合国家相关资质要求。并协调地方卫生部门做好检查工作。	立即组织对中毒人员进行救治，必要时送往最近医院进行检查治疗。同时对可疑的食品、饮水及其有关原料、工具设备和场所以及可能受污染的区域采取保留、控制措施，组织开展现场调查，迅速查明原因，并及时向大赛执委会报告。
设备损坏	提前一天服务器全部运行;现场划分备份组。	参赛选手举手示意后，监考人员计时，裁判确认后更换备机，并由主裁判确定应计入延时时间。
设备掉电	竞赛前技术人员及监考人员检查所有电源插头,确保牢固;电源线尽量绑扎在参赛选手碰不到的地方,如桌	参赛选手举手示意后，监考人员计时，裁判确认后重启机器，并由主裁判确定应计入延时的时间。

	子后面等。 竞赛前提醒参赛选手注意尽量不要碰到电源,配置文件要随时保存。	
现场网络线缆故障	现场走线要规范,尽量走暗槽或现场人员接触不到的地方;对主要线路要在走线槽内留有备线。	启用备线。
临时停电	赛场需要双路供电和备用发电机,确保单电源故障不会影响比赛	供电线路互为备份，如出现故障，切换线路，经裁判长与赛项执委会商议统一延长比赛时间；若双路电源均出现故障，快速启用备用发电机发电，保证比赛正常运行，经裁判长与赛项执委会商议统一延长比赛相应时间。

（二）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，可取消其继续比赛的资格。

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。

十四、赛项安全

赛项安全是全国职业院校技能大赛一切工作顺利开展的先决条件，是本赛项筹备和运行工作必须考虑的核心问题。

（一）组织机构

1.成立由赛项执委会主任为组长的赛项安全保障小组，成员包括承办院校主抓安全的校领导、学生工作处、后勤处、保卫处、合作企业技术工程师等相关人员；

2.与地方行政、交通、司法、安全、消防、卫生、食品、质检等相关部门建立协调机制，制定应急预案，及时处置突发事件，保证比赛安全进行。

（二）比赛环境

1.执委会须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。如有必要，也可进行赛场仿真模拟测试，以发现可能出现的问题。承办单位赛前须按照执委会要求排除安全隐患；

2.严格控制与参赛无关的易燃易爆以及各类危险品进入比赛场地，不许随便携带书包进入赛场；

3.配备先进的仪器，防止有人利用电磁波干扰比赛秩序。大赛现场需对赛场进行网络安全控制，以免场内外信息交互，充分体现大赛的严肃、公平和公正性；

4.大赛期间，承办单位须在赛场管理的关键岗位，增加力量，建立安全管理日志，在赛场封闭后至竞赛结束前对所有比赛场地进行监控，并将监控视频保留 3 个月，防止人为损坏大赛设备影响比赛正常进行。

（三）生活条件

1. 比赛期间，原则上由执委会统一安排参赛选手和指导教师食宿。承办单位须尊重少数民族的信仰及文化，根据国家相关的民族政策，安排好少数民族选手和教师的饮食起居;

2. 比赛期间安排的住宿地应具有宾馆/住宿经营许可资质。以学校宿舍作为住宿地的，大赛期间的住宿、卫生、饮食安全等由执委会和提供宿舍的学校共同负责;

3. 各赛项的安全管理，除了可以采取必要的安全隔离措施外，应严格遵守国家相关法律法规，保护个人隐私和人身自由;

4. 赛项所有裁判与参赛队住宿须在不同酒店。在竞赛日当天早 8 点，由竞赛执委会工作人员收缴裁判所有通信设备，直至竞赛成绩发布后再归还裁判;

5. 竞赛期间，除现场裁判外，其余裁判由竞赛执委会统一安排休息场所。在此期间，裁判人员不得随意出入，避免与参赛队代表取得联系。

（四）组队责任

1. 各学校组织代表队时，须安排为参赛选手购买大赛期间的人身意外伤害保险;

2. 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;

3. 各参赛队伍须加强对参与比赛人员的安全管理，实现与赛场安全管理的对接。

（五）应急处理

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。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，是否停赛由赛区执委会决定。事后，赛区执委会应向大赛执委会报告详细情况。

十五、竞赛须知

（一）参赛队须知

1.参赛队名称：统一使用规定的学校代表队名称，不使用其他组织、团体的名称；

2.参赛队组成：每支参赛队由 3 名参赛选手组成，须为同校在籍学生，其中队长 1 名。每支参赛队可配 2 名指导教师，指导教师须为本校专兼职教师。不接受跨校组队，同一学校报名参赛队不超过 1 支；

3.各参赛院校应指定 1 名负责人任赛项领队，全权负责该校参赛事务的组织、协调和领导工作；

4.参赛选手及指导教师在报名获得确认后，原则上不再更换。如在筹备过程中，参赛选手和指导教师因故不能参赛，须由其所在学校供职部门于赛项开赛前 10 个工作日之前出具书面说明，经大赛执委会办公室核实后予以更换。允许队员缺席比赛；允许指导教师缺席比赛；

5.参赛队按照大赛赛程安排，凭赛项执委会颁发的参赛证、有效身份证件和学生证参加比赛及相关活动；

6.赛项执委会统一安排各参赛队在比赛前一天进入赛场熟悉环境和设施情况；

7.参赛队选手、领队和指导教师要有良好的职业道德，严格遵守比赛规则和比赛纪律，服从裁判，尊重裁判和赛场工作人员，自觉维护赛场秩序；

8.领队应负责赛事活动期间本队所有选手的人身及财产安全，如

发现意外事故，应及时向赛项执委会报告;

9.各学校组织代表队时，须为参赛选手购买大赛期间的人身意外伤害保险;

10.对于有碍比赛公正和比赛正常进行的参赛队，视其情节轻重，按照《全国职业院校技能大赛奖惩办法》给予警告、取消比赛成绩、通报批评等处理。其中，对于比赛过程及有关活动造成重大影响的，以适当方式通告参赛院校或其所属地区的教育行政主管部门依据有关规定给予行政或纪律处分，同时停止该院校参加全国职业院校技能大赛1年。涉及刑事犯罪的移交司法机关处理。

（二）指导教师须知

1.严格遵守赛场的各项规定，服从裁判，文明竞赛。如发现弄虚作假者，取消参赛资格，名次无效;

2.领队和指导教师务必带好有效身份证件，在活动过程中佩戴“指导教师证”参加竞赛相关活动;

3.各代表队领队要坚决执行竞赛的各项规定，加强对参赛人员的管理，做好赛前准备工作，督促选手带好证件等竞赛相关材料;

4.在比赛期间要严格遵守比赛规则，不得私自接触裁判人员;

5.竞赛过程中，未经裁判许可，领队、指导教师及其他人员一律不得进入竞赛现场;

6.如对竞赛过程有疑议，由领队和指导教师负责以书面形式向大赛监督仲裁组反映，但不得影响竞赛进行;

7.对申诉的仲裁结果，领队要带头服从和执行，并做好选手工作。参赛选手不得因申诉或对处理意见不服而停止竞赛，否则以弃权处理;

8.领队和指导老师应及时查看有关赛项的通知和内容，认真研究

和掌握本赛项竞赛的规程、技术规范和赛场要求，指导选手做好赛前的一切技术准备和竞赛准备。

（三）参赛选手须知

1.参赛选手应严格遵守赛场规章、操作规程和工艺准则，保证人身及设备安全，接受裁判员的监督和警示，文明竞赛;

2.参赛选手应按照规定时间抵达赛场，凭身份证、学生证，以及统一发放的参赛证，完成入场检录、抽签确定竞赛赛位号，不得迟到早退;

3.参赛选手凭竞赛赛位号进入赛场，不允许携带任何电子设备及其他资料、用品;

4.参赛选手应在规定的时间段进入赛场，认真核对竞赛赛位号，在指定位置就座;

5.参赛选手入场后，迅速确认竞赛环境状况，填写相关确认文件，并由参赛队长确认签字（竞赛赛位号）;

6.参赛选手在收到开赛信号前不得启动操作。在竞赛过程中，确因计算机软件或硬件故障，致使操作无法继续的，经项目裁判长确认，予以启用备用计算机;

7.赛项任务书及相关资料，均保存在竞赛环境的“大赛资料”文件夹中。参赛选手应在竞赛规定时间内完成任务书内容，并按照要求，将相应文档按要求进行提交;

8.参赛选手需及时保存竞赛内容。对于因各种原因造成的数据丢失，由参赛选手自行负责;

9.参赛队所提交的答卷已在竞赛自动评分系统中通过赛位号标识，不得出现地名、校名、姓名、参赛证编号等信息，否则取消竞赛成绩;

10.竞赛过程中，因严重操作失误或安全事故不能进行比赛的(例如因操作原因发生短路导致赛场断电的、造成设备不能正常工作的)，现场裁判员有权中止该队比赛;

11.在比赛中如遇非人为因素造成的设备故障，经裁判确认后，可向裁判长申请补足排除故障的时间;

12.参赛选手不得因各种原因提前结束比赛。如确因不可抗因素需要离开赛场的，须向现场裁判员举手示意，经裁判员许可并完成记录后，方可离开。凡在竞赛期间内提前离开的选手，不得返回赛场;

13.竞赛时间结束，选手应全体起立，停止操作。将资料和工具整齐摆放在操作平台上，经工作人员清点后可离开赛场，离开赛场时不得带走任何资料;

14.在竞赛期间，未经执委会批准，参赛选手不得接受其他单位和个人进行的与竞赛内容相关的采访。参赛选手不得将竞赛的相关信息私自公布;

15.竞赛操作结束后，参赛队要确认成功提交竞赛要求的文件，裁判员在比赛结果的规定位置做标记，并与参赛队一起签字确认;

16.符合下列情形之一的参赛选手，经裁判组裁定后中止其竞赛:

（1）不服从裁判员/监考员管理、扰乱赛场秩序、干扰其他参赛选手比赛，裁判员应提出警告，二次警告后无效，或情节特别严重，造成竞赛中止的，经裁判长确认，中止比赛，并取消竞赛资格和竞赛成绩;

（2）竞赛过程中，由于选手人为造成计算机、仪器设备及工具等严重损坏，负责赔偿其损失，并由裁判组裁定其竞赛结束与否、是否保留竞赛资格、是否累计其有效竞赛成绩;

（3）竞赛过程中，产生重大安全事故、或有产生重大安全事

故隐患，经裁判员提示没有采取措施的，裁判员可暂停其竞赛，由裁判组裁定其竞赛结束，保留竞赛资格和有效竞赛成绩。

（四）工作人员须知

1.赛项全体工作人员必须服从执委会统一指挥，要以高度负责的态度做好比赛服务工作；

2.全体工作人员由赛项执委会统一聘用并进行工作分工，进入竞赛现场须佩戴赛项执委会统一提供的胸牌；

3.全体工作人员必须佩戴标志，认真检查证件，经核对无误后方可允许相关人员进入指定地点；

4.如遇突发事件要及时向执委会报告，同时做好疏导工作，避免重大事故发生，确保大赛圆满成功；

5.各工作组负责人，要坚守岗位，组织落实本组成员高效率完成各自工作任务，做好监督协调工作；

6.全体工作人员不得在比赛场内接打电话，以保证赛场设施的正常工作。

十六、申诉与仲裁

1.参赛队对不符合竞赛规定的设备、工具、软件，有失公正的评判、奖励，以及对工作人员的违规行为等，均可提出申诉；

2.申诉应在竞赛结束后2小时内提出，超过时效不予受理。申诉时，应按照规定的程序由参赛队领队向赛项监督仲裁工作组递交书面申诉报告。报告应对申诉事件的现象、发生的时间、涉及到的人员、申诉依据与理由等进行充分、实事求是的叙述。事实依据不充分、仅凭主观臆断的申诉将不予受理。申诉报告须有申诉的参赛选手、领队签名；

3.赛项监督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）领队向赛区监督仲裁委员会提出申诉。赛区监督仲裁委员会的仲裁结果为最终结果;

4.申诉人不得采取过激行为刁难、攻击工作人员，否则视为放弃申诉;

5.申诉方可随时提出放弃申诉。

十七、竞赛观摩

本赛项应须提供公开观摩区，使用大屏幕实时转播现场实况。

竞赛环境依据竞赛需求和职业特点设计，在竞赛不被干扰的前提下安全开放部分赛场。现场观摩应遵守如下纪律:

1.观摩人员需由赛项执委会批准，佩戴观摩证件在工作人员带领下沿指定路线、在指定区域内到现场观赛;

2.文明观赛，不得大声喧哗，服从赛场工作人员的指挥，杜绝各种违反赛场秩序的不文明行为;

3.观摩人员不得进入比赛区域，不可接触设备，同参赛选手、裁判交流，不得传递信息，不得采录竞赛现场数据资料，不得影响比赛的正常进行;

4.观摩者不可携带手机、IPAD等通讯工具进入赛场，对于各种违反赛场秩序的不文明行为，工作人员有权予以提醒、制止。

十八、竞赛直播

本赛项竞赛时采用全过程录像，在不影响比赛的前提下，全过程、全方位安排现场直播，并设直播观摩区，让所有参赛教师和社会人员等观看比赛。赛后邀请媒体采访优秀选手、优秀指导教师、裁判专家或企业人士，突出赛项的技能重点与优势特色，为大赛宣传、资源转

化提供全面的信息资料。视频资料也作为竞赛成果提交赛项执委会，作为竞赛历史材料供后续赛项提高进行参考，竞赛过程可作为教学资料进行资源转换，促进相关专业教学发展。

十九、资源转化

2021 年全国职业院校技能大赛大数据技术与应用赛项资源转化工作主要聚焦完善、升级已经开发完成的专业核心课程教学资源包，进一步开展师资培养，创新培训课程内容，建设大数据技术及其相关专业的生产实际教学案例库等工作，同时对产教融合校企合作案例进行总结。

二十、其他

无

附件：样卷

背景描述

企业消费服务平台，为大中小型企业提供基于云化的消费场景一站式智能消费、智能管控，帮助企业获得更高效、简单、美好的消费管理。从“费控+支付”出发，到覆盖全场景支出的创新模式，让员工在数字化平台上直接完成所有消费，从员工下单、到财务入账，全流程实现自动化统一结算、统一数据分析。解决传统差旅系统面临的场景覆盖不全、员工体验差、消费体验割裂等情况，真正做成一套让企业节省支出，让员工满意的差旅平台。

企业消费服务平台的出现将原来传统的差旅行程放到网络平台上,更广泛的传递差旅信息,互动式的交流更方便客人的咨询和订购,越来越多的人在出行的时候使用企业消费服务平台预订机票、火车票、住宿等，使得更多的商家愿意与企业消费服务平台建立合作，提升住宿场所的营业额，这也为企业消费服务平台的发展带来新的机遇，为了抓住这个机会，“企业消费服务平台”需要从地域、订单来源等多种维度进行分析，明确未来重点拓展合作商家的方向。公司要求多个小组进行分析，并提出相应建议，你所在的小组也在其中，需要通过数据采集、数据清洗、数据分析和数据可视化获得相关论据，提出未来重点拓展合作住宿场所的方向。

你们作为该小组的技术人员，是这次技术方案的核心成员，请按照下面步骤完成本次技术展示任务,并提交分析报告,祝你们成功!!!

模块 A: Hadoop 平台及组件的部署管理（15 分）

环境要求:

编号	主机名	类型	用户	密码
1	master	主节点	root	passwd
2	slave1	从节点	root	passwd
3	slave2	从节点	root	passwd

master01-1 主机上 MySQL 数据库用户名密码是 root/Password123$

相关软件安装包在/chinaskills 目录下

任务一: Hadoop 伪分布部署

本环节需要使用 root 用户完成相关配置，安装 Hadoop 需要配置前置环境，具体部署要求如下:

1、 解压 JDK 安装包到“/usr/local/src”路径，并配置环境变量，将命令（使用绝对路径）及环境变量内容复制粘贴至对应报告中;

2、 环境中已创建 ssh 密钥，实现主节点与从节点的无密码登录;截取主节点登录其中一个从节点的结果，将命令和结果复制粘贴至对应报告中;

3、 根据要求修改每台主机 host 文件，将 hosts 配置文件内容复制粘贴至对应报告中;

4、 在主节点修改 Hadoop 环境变量，并将（/etc/profile）配置文件内容复制粘贴至对应报告中;

5、 根据要求修改 Hadoop 相关文件（hadoop-env.sh、core-site.xml、HDFS-site.xml、mapred-site.xml、yarn-site.xml），初始化 Hadoop，并将初始化结果内容复制粘贴至对应报告中;

6、启动 Hadoop，使用相关命令查看所有节点 Hadoop 进程，并将结果内容复制粘贴至对应报告中。

任务二：Hive 组件部署

本环节需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体部署要求如下:

1、 解压 Hive 安装包到“/usr/local/src”路径，并使用相关命令，修改解压后文件夹名为 Hive，进入 Hive 文件夹，并将查看内容复制粘贴至对应报告中;

2、 配置 Hive 环境变量，并使环境变量只对当前用户生效，将环境变量内容复制粘贴至对应报告中;

3、 新建并配置 hive-site.xml 文件，实现“Hive 元存储”的存储位置为 MySQL 数据库，并将 hive-site.xml 配置文件内容复制粘贴至对应报告中;

4、 初始化 Hive 元数据（将 MySQL 数据库 JDBC 驱动拷贝到 Hive 安装目录的 lib 下），并将初始化结果内容复制粘贴至对应报告中;

5、 启动 Hive，检查是否安装成功，并将结果内容复制粘贴至对应报告中。

任务三、Kafka 组件部署

本环节需要使用 root 用户完成相关配置，已安装 Hadoop 及需要配置前置环境，具体部署要求如下:

1、 将 Zookeeper 配置完毕后，在各节点启动 Zookeeper，查看

Zookeeper 状态，并将命令和 Zookeeper 运行状态结果复制粘贴至对应报告中;

2、 修改 Kafkaserver.properties 文件，并将修改的内容复制粘贴至对应报告中;

3、 启动 Kafka，并将 Kafka 启动命令和输出结果前 10 行复制粘贴至报告中。

模块 B: 数据采集与处理（20 分）

1、 网站解析，利用 Chrome 查看网页源码，分析企业消费平台网站网页结构。

1） 打开企业消费平台网站，在网页中右键点击检查，或者F12快捷键，查看元素页面;

2） 检查网站: 浏览网站源码查看所需内容。

2、 从企业消费平台网站中爬取需要数据，按照要求使用 Python 语言编写爬虫代码，爬取指定数据项，并对结果数据集进行数据探索、以及必要的数据处理操作。请将符合题目要求的代码答案复制粘贴至对应报告中。

具体步骤如下:

1） 创建爬虫项目

2） 构建爬虫请求

3） 按要求定义相关字段

4） 获取有效数据

5） 将爬取到的数据保存到指定位置

至此已从住宿场所网站中爬取了所需数据，下一步我们要将爬取结果进一步进行相关数据操作。

详细数据描述:

1）请创建Scrapy项目chinaskills_accommodation（C:\chinaskills_accommodation），从网站（网站地址在竞赛平台模块B中给出）中爬取页面相关字段(包括name, seq, 业务部门, 拒

单率是否小于等于直销城市均值、，是否为客栈，房间价格，用户点评数，省份，酒店实住订单，酒店实住间夜）；将抓取结果保存为json格式文件，并命名为accommodations.json。每条信息请以Key: Value格式单独保存为一行数据。

例如:

{“name”:” ***”, ” seq”:” ***”,…….}

……

任务中要求将“以下内容及答案完整复制粘贴至对应报告中。”，粘贴到对应报告中的内容举例如下:

“中国”网页源码对应字段为: Country

“四川”网页源码对应字段为: Province

2）爬取数据量不少于28万条。

具体任务要求:

任务一: 网页源码应字段

使用 Chrome 浏览器，查找网站异步请求的数据，并将以下内容及答案完整复制粘贴至对应报告中。

“城市平均实住间夜”网页源码对应字段为:

“房间数”网页源码对应字段为:

“城市直销拒单率”网页源码对应字段为:

“处于商圈”网页源码对应字段为:

任务二、自行创建 Scrapy 工程

自行创建 Scrapy 工程编写爬虫代码，爬取“name、seq、业务部

门有效数据项包括：业务部门,房间数,国家,图片数,城市,城市平均实住间夜,城市直销拒单率,处于商圈”页面相关数据，通过爬虫代码分页爬取，以合理的程序逻辑判断相关数据包含的页数并将程序代码复制粘贴至对应报告中。

任务三：在 MySQL 中创建数据库表

根据爬取字段，在 MySQL 中创建 crawl 数据库，在该数据库中创建 accommodations1 表（包含 name,seq,业务部门,拒单率是否小于等于直销城市均值,是否为客栈,房间价格），创建 accommodations2 表（包含 name，seq，业务部门,用户点评数，省份，酒店实住订单,酒店实住间夜），将爬取数据写入相应数据表中，并分别统计 accommodations1 表和 accommodations2 表的总行数，将统计结果复制粘贴至对应报告中。

任务四：对数据库表排序

爬虫程序运行结束后查看 MySQL 数据库 accommodations1 表，按 seq 倒序排序，返回前 100 行数据，将命令与查看结果复制粘贴至对应报告中。

任务五：对数据表填充处理

请根据步骤 3 中 accommodations1 表中的数据，对数据集中“房间价格”字段的缺失值，使用平均值进行填充。查看填充后的数据集前 5 条记录，将查看结果复制粘贴至对应报告中。

任务六：对数据表删除处理

请根据步骤 3 中 accommodations2 表中的数据，对数据集中存在

空值的记录进行删除。查看删除后的数据集条数，将查看结果复制粘贴至对应报告中。

accommodations2 表删除后条数为: _____

模块 C：数据清洗与挖掘分析（25 分）

现已从相关网站及平台获取到原始数据集，为保障用户隐私和行业敏感信息，已进行数据脱敏。数据脱敏是指对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。在涉及客户安全数据或者一些商业性敏感数据的情况、不违反系统规则条件下，对真实数据进行改造并提供测试使用，如身份证号、手机号等个人信息都需要进行数据脱敏。

相关数据文件中已经包含了数据采集阶段从企业消费平台网站上爬取的数据集，其中包含了来自不同城市的多家住宿场所的销售信息，你的小组需要通过编写代码或脚本完成对相关数据文件中住宿场所销售管理数据的清洗和整理，并完成数据计算和分析任务。综合利用 MapReduce、Spark、Storm、分布式存储系统、数据仓库 Hive、数据推送工具等技术，使用 Java、Python、Scala 等开发语言，完成本阶段数据清洗、处理、分析及数据挖掘等任务。通过多个维度分析住宿场所的销售信息，并以此评价住宿场所销售业绩、区域的游客接纳能力、接纳质量等指标。

初始数据集来自多个网站及平台系统，且为多次采集汇总结果，因此数据集中不可避免地存在一些脏数据，即源数据不在给定的范围内或对于实际业务毫无意义，或是数据格式非法，以及在源系统中存在不规范的编码和含糊的业务逻辑。

请分析相关数据集，根据题目规定要求实现数据清洗及分析。

任务一、数据清洗

住宿场所销售数据涉及到多个平台及数据库对接，个别信息由于人为操作失误或计算机故障等原因产生了数据缺失值。缺失值是一种常见的脏数据情况，由于粗糙数据中缺少信息而造成的数据缺失或截断。现有数据集中某个或某些属性的值是不完全的。对于缺失值的处理，从总体上来说分为缺失值删除和缺失值插补。当缺失值过多时，信息条目本身的价值也会随之降低，此时如果对缺失值进行填补则将产生结果的人为干预。结合行业数据本身特点及上述考虑，请你根据题目具体参数要求实现以下功能：将缺失值大于 n 个的数据条目从原始数据集中剔除，并输出剔除的条目数量。

详细描述：

数据源文件存放于/chinaskills/accommodationdata.csv，请编写 MapReduce 程序，按照如下要求实现对数据的清洗，并将结果输出至 HDFS 文件系统中/accommodation_output1：

1） 解析该文件；

2） 按照题目要求剔除缺失数据信息（n=3），并以打印语句输出删除条目数；

3） 程序打包并在 Hadoop 平台运行，结果输出至 HDFS 文件系统中/ accommodation_output1。

具体任务要求：

1、 将 accommodationdata.csv 文件上传至 HDFS 新建目录 /file3_1 中；运行代码，删除数据源中缺失值大于 3 个字段的数

据记录，打印输出删除条目数，将运行结果复制粘贴至对应报告中;

2、 查看清洗后输出的结果文件总行数(/accommodation_output1)，将运行结果复制粘贴至对应报告中。

对于数据集字段缺失情况，通常可以采用填充默认值、均值、众数、KNN 填充、以及把缺失值作为新的 label 等方式处理。同时，不当的填充可能会令后续的分析结果出现导向性偏差，当缺失信息较少时可采用删除的方式来进行处理。下面请根据题目具体参数要求处理关键字段缺失，复制粘贴至对应报告中结果。

详细描述:

数据源使用 HDFS 文件系统中的 accommodationdata.csv，请编写 MapReduce 程序，按照如下要求实现对数据的清洗，并将结果输出至 HDFS 文件系统中/accommodation_output2:

1） 解析该文件;

2） 将任意关键字段为空的条目剔除，关键字段定义为{星级、评论数、评分}，并以打印语句输出删除条目数;

3） 程序打包并在 Hadoop 平台运行，结果输出至 HDFS 文件系统中/accommodation_output2。

具体任务要求:

3、 运行代码，将字段{星级、评论数、评分}中任意字段为空的数据删除，并打印输出删除条目数，将运行结果复制粘贴至对应报告中;

4、 查看清洗后输出的结果文件(accommodation_output2)总行数，将运行结果复制粘贴至对应报告中。

任务二、数据挖掘分析

城市游客接纳能力是城市规划建设中的重要指标，其中城市的住宿场所数量和房间数量是城市游客接纳能力的关键要素。请编写程序或脚本根据住宿场所管理网站中的数据统计各城市的相关信息，并写入指定的数据库或数据文件。

详细描述:

请根据数据清洗的输出数据集，编写 HQL 语句统计各城市的酒店出租率，以各城市酒店出租率降序排列并输出前 10 条统计结果，同时创建并写入数据表 a_4。要求输出字段包含：省份、城市、酒店出租率。

数据定义如下:

数据项	字段名	备注
省份	province	
城市	city	
酒店出租率	lease	要求保留 6 位小数

数据样式如下:

province	city	lease
贵州	贵阳	0.123456

具体任务要求:

1、 创建表 table3_4;

2、 统计各城市酒店出租率，将出租率前 10 的数据降序排列并写入数据表 table3_4 中，将命令复制粘贴至对应报告中。

企业消费平台是酒店营销的主要途径之一，不仅降低销售成本，同时也提高了顾客体验满意度。当顾客通过企业消费平台进行酒店预订时，酒店就拥有了用户的相关数据。通过这些数据，能够更好地收集用户需求，从而可以提供更有针对性和个性化的服务，最终能够产生更多的忠诚会员并带来更多订单。但企业消费平台销售也存在用户拒单等情况，拒单原因有很多：例如，平台信息不同步，信息更新不及时；分销层次过多，导致无法及时查证订单；酒店违反企业消费规则擅自以低价让客户取消订单，这种情况又叫做“切单”。企业消费平台需要统计用户订单的分布情况，以此发现平台缺陷及用户、商家的行为模式，企业消费平台据此调整营销策略。根据现有数据及给定参数完成订单数据统计，并写入指定的数据库或数据文件，复制粘贴至对应报告中结果。

详细描述：

请根据数据清洗的输出数据集，编写 HQL 语句统计各省直销拒单率，以直销拒单率升序排列并输出前 10 条统计结果，同时创建并写入数据表 table3_5。要求输出字段包含：省份、直销拒单率。

数据定义如下：

数据项	字段名	备注
省份	province	
直销拒单率	norate	要求保留 6 位小数

数据样式如下：

province	norate
贵州	0.123456

具体任务要求：

3、 创建表 table3_5，将命令复制粘贴至对应报告中；

4、 统计各省拒单率，将统计的拒单率升序排列并将前 10 条统计结果写入数据表 table3_5 中，将命令复制粘贴至对应报告中。

模块 D: 数据可视化（20 分）

MySQL 数据库中的相关数据集包含了城市、省份、评分、评论数等多项基础信息字段。请使用 Flask 框架，结合 Echarts 完成下列题目。

数据库账号: takeout　　密码: takeout

自行创建代码工程路径为: C:\chinaskills_hotel

每个可视化图中需要添加图片作为背景水印。

任务一: 柱状图呈现城市出租率

出租率是反映住宿场所经营状况的一项重要指标，它是已出租的客房数与住宿场所可以提供租用的房间总数的百分比。住宿场所出租率的情况可以在一定程度上反应出该住宿场所的整体运营的情况，为了更好的分析指定住宿场所的入住情况，请根据相关表中数据完成出租率分析，通过指定图例进行呈现。

详细描述:

请以数据库相关表作格为数据源，以柱状图呈现城市出租率。

具体任务要求:

1）　提取表格相关字段，在控制台按照“各省住宿场所出租率”降序排列，打印输出各省名称及包含的住宿场所数量;

打印语句格式如下:

==1.***省=住宿场所数为***个=出租率为: ***===

==2.***省=住宿场所数为***个=出租率为: ***===

……

2） 使用 Flask 框架，结合 Echarts 绘制柱状图。主标题为“各省住宿场所出租率”（字体要求：红色、加粗、斜体），副标题为出租率前十的省份；纵坐标为出租率，横坐标为省份名称（按照出租率降序排列）；将可视化结果复制粘贴至对应报告中。

任务二：折线图呈现连锁住宿场所出租率

连锁住宿场所一般都具有全国统一的品牌形象识别系统、全国统一的会员体系和营销体系、价格相比较很有优势，更适合大众化消费。连锁住宿场所无论在装修、服务还是信誉上都有较大的竞争优势，所以连锁住宿场所是出差、旅游住宿的首选。但是由于三线城市会员流动差、高素质管理人员相对短缺、营销环境与消费特点存在差异等问题，一些已经成熟住宿场所管理模式在三线城市可能并不受用，甚至会出现水土不服的现象。请根据现有数据及给定参数，统计指定连锁住宿场所的经营状况，并以指定图例进行呈现。

详细描述：

1） 数据库中相关表格已保存了指定地区的某连锁住宿场所销售信息。请根据地区划分，统计题中某连锁住宿场所的出租率（保留 6 位小数），并以折线图呈现；

2） 要求统计以下指定地区住宿场所相关信息，指定地区包括：东北、华北、华东、华中、西北、西南、华南；

3） 指定地区省份映射表，如表 1。

表 1：地区省份映射表

地区	省份
华东地区	山东、江苏、安徽、浙江、江西、福建、上海

华南地区	广东、广西、海南
华中地区	湖北、湖南、河南
华北地区	北京、天津、河北、山西、内蒙古
西北地区	宁夏、新疆、青海、陕西、甘肃
西南地区	四川、云南、贵州、西藏、重庆
东北地区	辽宁、吉林、黑龙江

具体任务要求:

1） 根据表格相关字段分别统计某连锁住宿场所在各地区的出租率（保留 6 位小数），在控制台按照“出租率”降序排列，打印输出各地区名称以及出租率;

打印语句格式如下:

==1.***地区，出租率为******

==2.***地区，出租率为******

……

2） 使用 Flask 框架，结合 Echarts 绘制折线图，主标题为“指定地区的住宿场所出租率“（字体要求：红色、加粗、斜体），副标题为“某连锁住宿场所的出租率”，纵坐标为出租率，横坐标为地区；输出折线图，将可视化结果复制粘贴至对应报告中。

任务三：散点地图呈现各城市住宿场所间夜数

住宿场所的间夜量也叫间夜数，是住宿场所在某个时间段内，房间出租率的计算单位，关于住宿场所间夜量的计算公式为间夜量=入住房间数*入住天数。例如某住宿场所今天入住的房间数为 500，则今天的间夜量=500*1=500，而又比如某住宿场所这个月（30 天）的平均每天入住房间数为 400，则这个月的间夜量=400*1*30=12000。

请根据指定表中数据统计住宿场所间夜数相关数据，并以指定图例进行呈现。

详细描述:

请以数据库相关表格中相关表作为数据源，各城市住宿场所间夜数散点地图。

具体任务要求:

1） 根据表格相关字段分别统计各城市住宿场所间夜数，打印输出各城市的间夜数，在控制台按照“间夜数”降序排列，打印输出各地区名称以及间夜数;

打印语句格式如下:

==***市: 间夜数为******

==***市: 间夜数为******

……

2） 使用 Flask 框架，结合 Echarts 绘制散点地图，标题为“各城市住宿场所间夜数”（字体要求: 红色、加粗、斜体）; 输出各城市住宿场所间夜数散点地图，将可视化结果复制粘贴至对应报告中。

任务四: 堆叠柱状图呈现直销和分销

订单数据是考量企业消费平台直销住宿场所经营业绩的重要指标，由于某些酒店资源无法内部消化，也会出现订单分销至其它企业消费平台的情况，此时称为分销。一般情况下，直销和分销是同时存在的。但当某些住宿场所或区域分销数量过多时，则表明企业消费平

台经营推广能力不足。请根据指定表中数据，以指定图例进行呈现。

详细描述:

根据相关负责人反馈，以下住宿场所的分销数量占比较大：山水时尚酒店北京梨园店，北京大宝饭店，北京普乐门白领公寓 798 精品店，北京长得福宾馆，北京中联鑫华酒店西客站店，北京瑞祥居宾馆，北京花神假日酒店。请使用数据库中相关数据，以堆叠柱状图呈现直销和分销，并辅以分销比率折线说明平台应对哪些酒店加强维护及推广力度。

具体任务要求:

1） 根据表格相关字段分别统计以上各家酒店的直销订单数量、分销订单数量以及分销比例，在控制台按照“分销比例”升序排列，打印输出各地区名称以及间夜数;

打印语句格式如下:

==“酒店名称：***直销订单数：***分销订单数：***分销比例:***”==

……

2） 使用 Flask 框架，结合 Echarts 绘制堆叠柱状图，并辅以分销比率折线，标题为“酒店直销，分销订单及比率”(字体要求：红色、加粗、斜体)；横坐标为酒店名称，纵坐标为销售数量和分销比例，将可视化结果截图并保存。

任务五：使用 sklearn 库中方法构建线性回归模型

企业消费平台为了能在更多省份扩展业务，与更多酒店建立合作

关系，为了赢得更多酒店的合作，在合作谈判过程中会通过同区域、同等级销售情况对比，需要提供同类酒店相关经营数据。请根据指定表中数据，以指定图例进行呈现。企业消费平台希望与住宿场所 A 进行线上销售合作，需要制作一份销售预测报告来说明酒店将在平台收获的间夜预期。住宿场所 A 信息{广东省、广州市、北京路商圈、非客栈，评论数 100，房间数 200}

详细描述:

请以根据表格相关字段: 是否客栈、评论数、房间数为特征变量，构建线性回归模型，给出明年同期住宿场所 A 在本平台总间夜数的预期值。输出预测模型相关指标，同时给出预期结果。

具体任务要求:

1）请使用 sklearn 库中方法构建线性回归模型，并在控制台输出住宿场所 A 总间夜的预测值;

打印语句格式如下:

==“住宿场所 A 明年同期总间夜数预期值为: ***” ===

2）使用 Flask 框架，结合 Echarts 绘制散点线性回归图，标题为“住宿场所 A 总间夜数预测”(字体要求: 红色、加粗、斜体)，横坐标为时间，纵坐标为总间夜数，将可视化结果截图并保存。

任务六: 多线雷达图呈现各省份住宿场所综合情况

企业消费平台需要综合评判一个城市住宿场所运营情况，会涉及到多方面住宿场所数据，例如像高端住宿场所数量、订单数量、住客

评分、评论数量、出租率、200 元/晚以下快捷住宿场所数量等信息，请根据指定表中数据统计相关数据，并以指定图例进行呈现。

详细描述:

请根据数据库中相关表格，统计各城市住宿场所综合运营况，并以多线雷达图表达。

具体任务要求:

1）根据数据库中相关表格分别统计北京、上海、广东、四川、海南各地四星/五星住宿场所的数量、平均评分、评论数、各省住宿场所出租率、直销拒单率，在控制台按照“省份”名称升序排列，打印输出各城市住宿场所的多项运营指标;

打印语句格式如下:

==省市：A，四星/五星住宿场所数量为：******

==省市：A，平均评分为：******

……

==省市：B，四星/五星住宿场所数量为：******

==省市：B，平均评分为：******

……

2）使用 Flask 框架，结合 Echarts 绘制多线雷达图，标题为各省份住宿场所综合情况（字体要求：红色、加粗、斜体）；输出多线雷达图，将可视化结果复制粘贴至对应报告中。

模块 E：综合分析（20 分）

假定你为企业消费平台的管理者，在综合理解住宿场所业务数据的基础上，通过以上模块 A、B、C、D 的相关结论，对未来拓展合作住宿场所方向做出预测，根据题目要求进行分析，并编写输出分析报告。

根据上述任务中的结论，分析以下内容，并编写分析报告。从住宿场所分布维度，结合多省份住宿场所综合运营情况，对企业消费平台未来拓展合作住宿场所的方向提出建议。

分析报告要求：

任务一：通过数据及图示分析原因

结合平台相关数据文件，以各省住宿场所出租率和各城市住宿场所间夜数的折线图，对各省住宿场所的运营情况进行分析，分别以文字描述和图例进行说明；

任务二：对通过图示和计算业务分析原因

结合模块 D 可视化分析对某连锁酒店在不同地区的酒店出租率的统计，说明影响酒店出租率的原因可能有哪些？对于提高该连锁酒店的出租率，您有哪些建议？分别以文字描述和图例进行说明；

任务三：对企业消费平台未来拓展合作建议和意见

对企业消费平台未来拓展合作住宿场所的方向提出建议（不少于 3 条建议）；

附录：补充说明

一、 json 数据格式样例

{"name": "南京国美家庭旅社公寓南林店", "detail": {"SEQ": "nanjing_10116", "国家": "中国", "省份": "江苏", "城市": "南京", "处于商圈": "锁金村地区 玄武湖地区 中山陵景区","是否为客栈": 0, "住宿场所星级": "二星及其他","业务部门": "低星", "剩余房间": 8, "图片数": 0, "住宿场所评分": "1", "用户点评数": 1,"城市平均实住间夜": "51.701686747", "住宿场所总订单": 0,"住宿场所总间夜": 0, "住宿场所实住订单": 0, "住宿场所实住间夜": 0, "住宿场所直销订单": 0, "住宿场所直销间夜": 0, "住宿场所直销实住订单":0, "住宿场所直销实住间夜": 0,"住宿场所直销拒单": 0, "住宿场所直销拒单率": null, "城市直销拒单率": "0.0282838180927","拒单率是否小于等于直销城市均值": 0, "最低房间价格": "306"}}

二、 fastjson-1.2.41.jar 常用 API（java）

1、 实例化

JSONObject ( );

2、 JSON 解析包

com.alibaba.fastjson.JSON;

com.alibaba.fastjson.JSONObject;

com.alibaba.fastjson.JSONArray;

com.alibaba.fastjson.JSONException;

3、 常用 API 方法:

1） public static final Object parse (String text); // 把 JSON 文本 parse 为 JSONObject 或者 JSONArray

2） public static final JSONObject parseObject (String text); // 把 JSON 文本 parse 成 JSONObject

3） public static final T parseObject (String text, Class clazz); // 把 JSON 文本 parse 为 JavaBean

4） public static final JSONArray parseArray (String text); // 把

JSON 文本 parse 成 JSONArray

5）public static final List parseArray（String text, Class clazz）; //把 JSON 文本 parse 成 JavaBean 集合

6）public static final String toJSONString（Object object）; // 将 JavaBean 序列化为 JSON 文本

7）public static final String toJSONString（Object object, boolean prettyFormat）; // 将 JavaBean 序列化为带格式的 JSON 文本

8）public static final Object toJSON(Object javaObject); 将 JavaBean 转换为 JSONObject 或者 JSONArray。

三、 fastjson-1.2.41.jar 常用 API【Spark（scala）】

1、 json 解析包

com.alibaba.fastjson.JSON

2、 常用 API

1）实例化:

JSON.parseObject（x）

2）默认值: 如果该 key 没有值默认为 null:

jsonObject.getOrDefault（key,默认值）

jsonObject.getOrDefault（"name",""）

3）获取该 key 的 value 值

jsonObject.get（json 的 key）

jsonObject.get（“name”）

4）判断 key 是否存在

jsonObject.containsKey（key）

5）添加 kv 键值对

jsonObject.put（key,value）

四、 控制台输出运行日志样例

五、 方差、均方根差的定义

1、 方差MSE: 概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。

2、 均方根差RMSE: 均方根误差，是观测值与真值偏差的平方和观测次数n比值的平方根。RMSE是计算观测值与其真值，或者观测值与其模拟值之间的偏差。

六、 间夜定义

间夜又称间夜数，是住宿场所在某个时间段内，房间出租率的计算单位。例如 20 间房入住 2 晚，为 40 间夜数。

七、 出租率计算公式

出租率 = 当月发生的总间夜数/当月所能提供的总房间数

八、 线性回归预测数据源 data_accommodation_mult.csv 字段名

SEQ、省份、城市、商圈、是否为客栈、星级、房间数、评论数、平均评分数、城市平均间夜、住宿场所总订单、住宿场所总间夜、住宿场所实住订单、住宿场所实住间夜、住宿场所直销订单、住宿场所直销实住订单、住宿场所直销间夜、住宿场所直销实住间夜、城市直销拒单、城市直销拒单率、住宿场所企业消费平台实住订单

九、 数据可视化表字段说明

表 radar_lines

province	省份
accommodation_num	住宿场所数
avg_score	平均分

comment_num	评论数
lease_rate	出租率
direno_rate	直销率

表 platform_rate

accommodationname	住宿场所名称
provice	省份
city	城市
trading_area	商圈
is_inn	是否为客栈
start	星级
room	房间数
commen	评论数
score	评分
city_avgorder	城市平均订单
city_avgmidnight	城市平均间夜
city_avgrealorder	城市平均实住订单
city_avgrealmidnight	城市平均实住间夜
accommodation_order	住宿场所订单
accommodation_midnight	住宿场所总间夜
accommodation_realorder	住宿场所实住订单
accommodation_realmidnight	住宿场所实住间夜
accommodation_direorder	住宿场所直销订单
accommodation_diremidnight	住宿场所直销间夜
accommodation_direrealmidnight	住宿场所直销实住间夜
accommodation_direnoorder	住宿场所直销拒单
city_direorder	城市直销订单
city_direrealorder	城市实住订单
city_direnoorderrate	城市直销拒单率
……	……..

表 platform

province	省份
city	城市
order_num	总订单
midnight_num	总间夜
real_num	实住订单
realmidnight_num	实住间夜
city_lease_rate	出租率

表 city_let_rate

province	省份
t 企业消费平台 l_orders	总订单
t 企业消费平台 l_nigth	总间夜
real_orders	实住订单
real_night	实住间夜
city_rate	出租率

表 night

province	省份
city	城市
night	间夜