## 模块A：大数据平台部署管理

本模块用到的所有安装包均在Master节点的/chinaskills目录下。

### Hadoop全分布式部署:

1. 将JDK包解压到/usr/local/src路径，将完整命令复制粘贴到对应报告中;
2. 修改/root/.bash\_profile文件，设置JDK环境变量，并使环境变量只对当前root用户生效将环境变量配置内容复制粘贴至对应报告中;
3. 从master复制上面步骤配置的JDK环境变量文件到slave1、slave2节点，命令和结果复制粘贴在对应报告中;
4. 配置SSH密钥登录，实现从master登录到slave1，将登录命令和结果复制粘贴在对应报告中;
5. 将配置文件hadoop-env.sh变更内容复制粘贴在对应报告中；
6. 将配置文件core-site.xml变更内容复制粘贴在对应报告中；
7. 初始化Hadoop环境namenode，将命令及结果复制粘贴在对应报告中；
8. 查看master及slave1节点jps进程，将其命令及结果复制粘贴至对应报告中。

### ZooKeeper部署管理:

1. 解压Zookeeper安装包到“/usr/local/src”路径，并修改解压后文件夹名为zookeeper，将修改命令及结果复制粘贴至对应报告中；
2. 设置ZOOKEEPER\_HOME环境变量，并使环境变量只对当前用户生效，命令及环境变量内容复制粘贴至对应报告中；
3. 配置“zoo.cfg”配置文件，将文件变更内容复制粘贴至对应报告中；
4. 修改myid配置文件，将文件变更内容复制粘贴至对应报告中；
5. 启动每个虚拟机上的Zookeeper节点，启动完成之后查看每个节点的zkServer服务状态，命令及结果复制粘贴至对应报告中。

### Kafka组件部署管理

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。

### Hive 部署管理

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Hive安装包解压到（/usr/local/src）下，使用绝对路径，将命令复制并粘贴至对应报告中；
2. 把解压后的apache-hive-1.1.0-bin文件夹更名为hive；进入hive文件夹，并将查看命令及结果复制并粘贴至对应报告中；
3. 设置Hive环境变量，并使环境变量只对当前root用户生效；并将环境变量配置内容复制并粘贴至对应报告中；
4. 将Hive安装目录里hive-default.xml.template文件更名为hive-site.xml；并将更改命令复制并粘贴至对应报告中；
5. 通过VI编辑器配置hive-site.xml文件，将MySQL数据库作为Hive元数据库。将配置文件Hive元存储相关内容复制并粘贴至对应报告中；
6. 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文件夹下；并通过schematool执行初始化，将初始化结果复制粘贴至对应报告中；
7. 启动Hive并保存命令输出结果，将结果输出复制粘贴至对应报告中。

### Spark(Standalone模式)组件部署管理

1. 解压scala安装包到“/usr/local/src”路径下，并更名为scala，命令及结果复制粘贴至对应报告中；
2. 设置scala环境变量，并使环境变量只对root用户生效，配置文件内容复制粘贴至对应报告中；
3. 进入scala命令行界面，将命令及结果复制粘贴至对应报告中；
4. 解压Spark安装包到“usr/local/src”路径下，并更名为spark，将命令及结果复制粘贴至对应报告中；
5. 设置Spark相关环境变量，并使环境变量只对root用户生效，配置Spark的master节点主机名、端口、worker结点的核数、内存，将命令复制粘贴至对应报告中；
6. 启动Spark shell，将命令及结果复制粘贴至对应报告中。

### Sqoop部署管理

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Sqoop安装包到/usr/local/src路径下，并使用相关命令，修改解压后文件夹名为sqoop，进入sqoop文件夹，并将查看内容复制粘贴至对应报告中；
2. 修改Sqoop环境变量，并使环境变量只对当前root用户生效,将环境变量配置内容复制粘贴至对应报告中；
3. 修改并配置sqoop-env.sh文件，将命令及结果复制粘贴至对应报告中；
4. 测试Sqoop连接MySQL数据库是否成功，将命令及结果复制粘贴至对应报告中。具体任务要求：使用sqoop命令连接MySQL数据库，查询MySQL中所有数据库名称，将命令及结果复制粘贴至对应报告中。

### Flume部署管理

1. 设置Flume环境变量，并使环境变量只对当前root用户生效，将变量内容复制并粘贴
2. 修改并配置flume-env.sh文件，将修改内容复制并粘贴到对应报告中。
3. 查看Flume启动的conf-file文件内容，将内容复制并粘贴到对应报告中；
4. 启动Flume传输Hadoop日志，将Flume启动命令以及查看HDFS上生成的Hadoop日志文件结果复制并粘贴到对应报告中（若文件数量大于10则只截取10条）。并查看HDFS中/tmp/flume目录下生成了文件，将内容复制并粘贴到对应报告中。

## 模块B：数据采集与处理

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。
2. 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；
3. 检查网站：浏览网站源码查看所需内容。
4. 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

至此已从企业消费平台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s\_accommodation（C:\ s\_accommodation），从网站中爬取页面相关字段（包括住宿场所订单, 住宿场所总间夜, 星级, 住宿场所直销订单, 住宿场所直销实住间夜等）；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”:”\*\*\*”,…….}

**具体任务要求：**

### 任务一：网页源码对应字段

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”网页源码对应的当前div的上层td的源代码为：

“房间数”网页源码对应的当前div的上层td的源代码为：

“城市直销拒单率”网页源码对应的当前div的上层td的源代码为：

“商圈”网页源码对应的当前div的上层td的源代码为：

“是否为客栈”网页源码对应的当前div的上层td的源代码为：

“住宿场所实住订单”网页源码对应的当前div的上层td的源代码为：

“住宿场所直销拒单” 网页源码对应的当前div的上层td的源代码为:

“住宿场所直销订单” 网页源码对应的当前div的上层td的源代码为：

“住宿场所直销间夜” 网页源码对应的当前div的上层td的源代码为:

“评分”网页源码对应的当前div的上层td的源代码为:

### 任务二：创建Scrapy工程

1. 自行创建Scrapy工程编写爬虫代码，通过使用Chrome浏览器开发者工具进行抓包，将Scrapy 默认UserAgent 配置为Chrome浏览器UA，并将配置代码复制粘贴至对应报告中。
2. 将Scrapy 默认请求头的 Accept和Accept-Language 设置为抓包到的值，并将配置代码复制粘贴至对应报告中。
3. 在配置文件中配置Scrapy 的下载延时为2秒，并将配置代码复制粘贴至对应报告中。
4. 爬取有效数据项包括：“住宿场所名称,城市,商圈,住宿场所订单,住宿场所总间夜,星级,住宿场所直销订单,住宿场所直销实住间夜”，通过爬虫代码分页爬取，将使用re解析分页链接的程序源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（包含住宿场所名称,城市,商圈,住宿场所订单,住宿场所总间夜），创建accommodations2表（包含住宿场所名称,城市,商圈,星级,住宿场所直销订单,住宿场所直销实住间夜）并查看表结构，将查看结果（含字段总行数）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accommodations2表，按住宿场所直销订单字段倒序排序，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

1. 请根据步骤3中accommodations2表中的数据，对数据集中“住宿场所直销实住间夜”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15条记录，将查看结果复制粘贴至对应报告中。
2. 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

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(MapReduce)

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/jd\_4706.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大于n个字段的记录（n=3），并以打印语句输出删除记录数；
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jd\_4706.csv文件上传至HDFS新建目录/file3\_1中；在Hadoop平台上运行代码，删除数据源中缺失值大于3个字段的数据记录，打印输出删除记录数，将运行结果复制粘贴至对应报告中；

打印语句格式如下：

=== “删除缺失值大于3个的字段的数据条数为\*\*\*条”===

1. 使用Hadoop Shell命令统计查看清洗后输出的结果文件(/accommodation\_output1)总行数，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jd\_4706.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记录剔除，关键字段定义为{星级、评论数、评分}，并以打印语句输出删除记录数；
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：**

1. 在Hadoop平台运行代码，将字段{星级、评论数、评分}中任意字段为空的数据删除，并打印输出删除记录数，将运行结果复制粘贴至对应报告中；

打印语句格式如下：

=== “删除的数据条数为\*\*\*条”===

1. 使用Hadoop Shell命令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：数据清洗(Spark)

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/jd\_4706.csv文件夹下，请编写Spark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/accommodationsparktask1。

1. 解析/chinaskills下相关数据文件；
2. 按照题目要求剔除缺失数据信息大于n（n=3）个字段的数据记录，并以打印语句输出删除条目数；
3. 程序打包并在Spark上运行，结果（包含csv的表头）输出至HDFS文件系统中/accommodationsparktask1。

**具体任务要求：**

1.编写Spark(Standalone)程序，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

打印语句格式如下：

=== “删除缺失值大于3个的字段的数据条数为\*\*\*条”===

2.使用Hadoop shell命令查看清洗后输出的结果文件/accommodationsparktask1总行数(包含csv的表头)，将查看命令与执行结果复制粘贴至对应报告中。

缺失值是一种常见的脏数据情况，对于数据缺失值的处理，从总体上来说分为缺失值删除和缺失值插补两种处理方式。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，则数据分析结果可能会受到干扰，有失客观性。结合行业数据本身特点及上述考虑，对于数据集中数值字段缺失的情况，通常可以采用填充固定值、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。

**详细描述：**

请以上述**1、（题目数字编号）**任务结果数据集作为输入数据源，编写Spark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/accommodationsparktask2。

1. 解析/accommodationsparktask1中的数据文件；
2. 将关键字段“星级”、“评论数”、“评分”中存在缺失值的数据删除；
3. 程序打包并在Spark(Standalone)集群上运行，结果（包含csv的表头）输出至HDFS文件系统中/accommodationsparktask2。

**具体任务要求：**

3.在Spark(Standalone)上运行代码，将字段“星级”、“评论数”、“评分”中存在缺失值的数据删除，打印输出删除条目数，将打印输出结果复制并保存至对应报告中（复制内容需包含打印语句输出结果的上下各5行运行日志）。

打印语句格式如下：

=== “删除的数据条数为\*\*\*条”===

4.使用Hadoop shell命令查看清洗后输出的结果文件/accommodationsparktask2总行数（包含csv的表头），将查看命令与执行结果复制粘贴至对应报告中。

### 任务三：数据挖掘分析-HQL实现出租率统计

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：**

1. 请根据数据清洗的输出数据集，先建立Hive外表external\_data\_table，编写Hive SQL语句统计各城市的住宿场所出租率，以各城市住宿场所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table3\_3。要求输出字段包含：省份、城市、住宿场所出租率。

数据定义如下：

|  |  |  |
| --- | --- | --- |
| 数据项 | 字段名 | 备注 |
| 省份 | province |  |
| 城市 | city |  |
| 住宿场所出租率 | 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. 在HIVE的defalut库中创建外表external\_data\_table，外表的文件存储在hdfs的/accommodation\_output2,将建表SQL复制粘贴至对应报告中；
2. 利用步骤1表external\_data\_table数据统计各城市住宿场所出租率，将出租率前10的数据降序排列并写入数据表在Hive的defalut库中的table3\_3中，将建表SQL、数据写入SQL以及数据查询结果复制粘贴至对应报告中。

### 任务四：数据挖掘分析-HQL实现直销拒单率统计

企业消费平台是住宿场所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住宿场所预订时，住宿场所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；住宿场所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：**

1. 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\_4。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：**

1. 在HIVE的defalut库中创建内表table3\_4，将建表SQL复制粘贴至对应报告中；
2. 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\_4中，将数据写入SQL以及数据查询结果复制粘贴至对应报告中。

### 任务五：数据挖掘分析-Spark实现场所数量和房间数量统计

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：**

请以数据清洗与挖掘分析模块任务一的输出结果作为输入数据源，编写Spark程序统计各城市的住宿场所数量和房间数量，以城市房间数量降序排列并输出前10条统计结果，输出至HDFS文件系统中/accommodationhive1，同时创建并写入数据表table3\_5。要求输出字段包含：省份、城市、住宿场所数量、房间数量。

**具体任务要求：**

1. 编写Spark程序，统计各城市的住宿场所数量和房间数量，以城市房间数量降序排列，输出前10条统计结果至HDFS文件系统中/accommodationhive1，使用Hadoop shell命令查看统计结果，将查看命令与执行结果复制并粘贴至对应报告中；
2. 在Master节点 的MySQL数据库（数据库账号：root，数据库密码：123456）中创建accommodationdata数据库并在其中创建表table3\_5，将统计结果写入表table3\_5中，输出表table3\_5前10行数据，将MySQL数据查询命令及数据查询结果复制并粘贴至对应报告中。

### 任务六：数据挖掘分析-MapReduce实现直销拒单率统计

通过收集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；住宿场所违反企业消费平台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。

**详细描述：**

1. 请根据数据清洗的输出数据集，编写MapReduce程序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\_6；
2. 要求输出字段包含：省份、直销拒单率；

**具体任务要求：**

1. 在accommodationdata数据库中创建表table3\_6，将建表命令复制并粘贴至对应报告中；
2. 运行代码，将统计的直销拒单率按升序排列，并将前10条统计结果输出至HDFS文件系统/accommodationhive2中，并写入数据表table3\_6中，将MySQL数据查询命令及数据查询结果复制并粘贴至对应报告中。

## 模块D：数据可视化

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：root 密码：123456 库名：hotel

自行创建代码工程路径为：C:\s\_hotel

### 任务一：条形图呈现住宿场所订单量

住宿场所订单量是反映住宿场所入住数量的重要指标之一，某省订单数量一定程度上可以反应出该省住宿场所入住情况，为了更好地分析全国各省住宿场所订单量，请根据指定表中数据统计出全国各省住宿场所订单量的情况，并以指定图例进行呈现。

**详细描述：**

请以数据库中相关表格为数据源，分析并统计各省订单数量，绘制条形图。

**具体任务要求：**

1. 提取表格相关字段，分别统计各省的订单数量，在控制台按照“订单数量”降序排列，打印输出各省份的订单数量；

打印语句格式如下：

==省份：\*\*\* =订单数量：\*\*个==

……

1. 使用Flask框架，结合Echarts绘制“订单数量”的条形图，标题为“各省订单数量”（字体颜色：红色，加粗），纵坐标为省份名称，横坐标为订单数量，并将可视化结果复制粘贴至对应报告中。

### 任务二：多个饼图呈现各省份不同等级住宿场所占比

企业消费平台为了更好地发展企业业务，向企业客户推荐符合其定位的协议住宿场所，需要分析上题中Top5省份的“三星级/舒适”、“四星级/高档”和“五星级/豪华”住宿场所以及“其它类别”住宿场所（除上述三类外，其余类型住宿场所均归为“其它类别”）的占比情况。请根据指定表中数据，以指定图例进行呈现。

**详细描述：**

请根据数据库中相关数据集中省份、星级等相关字段统计各省份不同等级住宿场所的占比，绘制多个饼图。

**具体任务要求：**

1. 根据表格相关字段分别统计上题Top5省份不同星级住宿场所的占比，打印输出各省份名称以及各星级住宿场所的占比情况；

打印语句格式如下：

==省份：A=舒适型住宿占比：\*\*===

==省份：A=高档型住宿占比：\*\*===

==省份：A=豪华型住宿占比：\*\*===

==省份：B=舒适型住宿占比：\*\*===

==省份：B=高档型住宿占比：\*\*===

==省份：B=豪华型住宿占比：\*\*===

……

1. 使用Flask框架，结合Echarts在组合图中绘制多个饼图，每个省份各星级住宿场所占比情况分别用一个饼图进行呈现，将可视化结果截图并保存。

### 任务三：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：**

请以数据库中相关表格为数据源，分析并统计各省住宿场所出租率，使用柱状图表达。

**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以及出租率；

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：\*\*\*===

……

1. 使用Flask框架，结合Echarts绘制柱状图。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）；将可视化结果截图并保存。

### 任务四：折线图呈现连锁住宿场所出租率

一些全国连锁的住宿场所品牌通常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符合大众化消费。连锁品牌无论在装修、服务还是信誉上都有较大的竞争优势，所以连锁品牌的住宿场所是出差、旅游住宿的好选择。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的差异等问题，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. 数据库中相关表已保存了指定地区的某连锁住宿场所的销售信息。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\*\*\*===

……

1. 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，将可视化结果截图并保存。

### 任务五：堆叠柱状图呈现各省住宿场所数和订单信息

在该企业消费平台上，各地区的住宿场所销售数据等信息能够反映一个地区商业活动的密集程度。例如住宿场所总量多的城市大都具有强烈的吸纳外来人员的能力，订单数量能够反映该地区的有较多的商业往来。根据现有数据及给定参数完成住宿场所销售数据统计。

**详细描述：**

请根据数据库中相关数据集中住宿场所数量、订单数量、实住数量等字段统计各省份的住宿场所数量和订单信息，绘制堆叠柱状图。

**具体任务要求：**

1. 提取表格相关字段，分别统计各省的住宿场所数量、总订单数量和实住订单数量，在控制台按照“住宿场所数量”降序排列，打印输出住宿场所数量最多的5个省份及其包含的住宿场所数量、订单数量和实住数量；

打印语句格式如下：

==省份：\*\*\*=住宿场所数量：\*\*个=订单数量：\*\*个=实住数量：\*\*个==

……

1. 使用Flask框架，结合Echarts，按照“总订单数量”降序排列，绘制“总订单数量”与“实住订单数量”的堆叠柱状图。标题为“住宿场所数量Top5”，横坐标为省份名称，纵轴分别表示住宿场所数量、订单数量和实住数量

### 任务六：堆叠柱状图呈现直销和分销直销和分销

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:\*\*\*”==

……

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。

### 任务七：玫瑰饼图呈现各地区高端场所占比

高端住宿场所的数量，能够从一个侧面反映当地的经济水平，据国家旅游和文化部统计境内31个省市（不含港澳台）共有860家五星级住宿场所，但分布很不均衡，其中东部沿海所占有的五星级数量，接近了全国一半。请你根据题目要求统计各地区的高端住宿场所数量及各地区高端住宿场所占全国所有高端住宿的比例。请按题目要求，输出相关图例。

**详细描述：**

1. 数据库中相关表已保存了各地区的高端住宿场所信息；
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各地区高端住宿场所（五星级）的数量，以及各地区高端住宿场所占全国住宿场所总量的比例，并按“高端住宿场所”降序进行排列；

打印语句格式如下：

==1.地区：\*\*\*, 高端住宿场所数量：\*\*\*，所占比例：\*\*\*===

==2.地区：\*\*\*, 高端住宿场所数量：\*\*\*，所占比例：\*\*\*===

……

使用Flask框架，结合Echarts根据各地区高端场所数量占全国总量比例绘制玫瑰图，标题为“各地区高端场所数量所占比例”（字体要求：红色、加粗、斜体），并标识各分类占比，输出玫瑰饼图，将可视化结果截图并保存。

### 任务八：条形统计图呈现各省直销拒单率

企业消费平台是住宿场所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住宿场所预订时，住宿场所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；住宿场所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。根据现有数据统计订单数据，并以指定图例进行呈现。

**详细描述：**

请根据数据库中相关数据集中统计各省直销拒单率，以直销拒单率降序排列，并绘制条形统计图。

**具体任务要求：**

1. 根据表格相关字段分别统计各省份的直销拒单率（保留6位小数），在控制台按照“直销拒单率”降序排列，打印输出各省名称以及直销拒单率；

打印语句格式如下：

==1.\*\*\*省, 直销拒单率为\*\*\*===

==2.\*\*\*省，直销拒单率为\*\*\*===

……

1. 使用Flask框架，结合Echarts绘制直销拒单率最高的5个省份条形图，主标题为“直销拒单率最高的5个省份”（字体要求：红色、加粗、斜体），纵坐标为省份名称，横坐标直销拒单率；输出条形图，将可视化结果截图并保存。

### 任务九：散点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\*\*\*===

……

使用Flask框架，结合Echarts绘制散点图，主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图，将可视化结果截图并保存。

### 任务十：多线雷达图呈现各省份住宿场所综合情况

企业销售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：**

请以数据库中相关表格为数据源，统计北京、上海、广东、四川、海南各省市住宿场所的“平均评分”、”评论数”、”各省住宿场所出租率”等综合指标，并以多线雷达图。

**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，平均评分为：\*\*\*===

==省市：A，评论数：\*\*\*===

……

==省市：B，平均评分为：\*\*\*===

==省市：B，评论数：\*\*\*===

……

1. 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，将可视化结果截图并保存。

# 附录：补充说明

## 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"}}

## -控制台输出运行日志样例



## 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。

## 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

## 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 |
| province | 省份 |
| city | 城市 |
| trading\_area | 商圈 |
| is\_in | 是否为客栈 |
| star | 星级 |
| room | 房间数 |
| comment | 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 |