练习题一

任务一：Hadoop 平台及组件的部署管理（15 分）

一、 Hadoop 全分布部署

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；

2、 在指定目录下安装 ssh 服务，查看 ssh 进程并截图（安装包统一在“/h3cu/”）；

3、 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；

4、 根据要求修改每台主机 host 主机名；

5、 修改每台主机 host 文件配置 IP 与主机名映射关系；

6、 根据要求修改 Hadoop 环境变量；

7、 根据要求修改 Hadoop 相关文件，并初始化 Hadoop；

8、 启动 Hadoop，使用相关命令查看所有节点 Hadoop 进程并截图。

二、 Flume 组件部署

1、 解压 Flume 安装包到“/usr/local/src”路径；

2、 修改解压后文件夹名为 flume；

3、 设置 Flume 环境变量，并使环境变量只对当前 root 用户生效；

4、 修改 Flume 相应文件；

5、 修改并配置 flume-env.sh 文件。

任务二：数据采集（15 分）

（自行搭建网站，参考源数据）

网站数据文件路径：/h3cu/mysql.excl

1、 网站解析，利用 chrome 查看网页源码，分析招聘网站网页结构。

1） 打开酒店网站，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、 从酒店网站中爬取需要数据，按照要求使用 Java 或 Python 语言编写并完善爬虫代码，爬取指定数据项，有效数据项包括但不限于：城市、商圈、星级、评分、评论数等多项字段。并将代码文件与代码截图保存。

具体步骤如下：

1） 创建爬虫项目\H3CU_hotel\

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

至此已从酒店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，请将操作命令截图并保存。

创建 scrapy 项目 ScrapyHotel。本任务要求从酒店网站中抓取数据，提取全部有效数据项。将爬取到的数据写入 Mysql 数据库中。

根据任务二题目要求，完成以下内容：

1、 通过对网站结构分析，编写并完成下表：

内容	标签
酒店编号	
酒店星级	
业务部门	
酒店评分	

2、 根据爬取字段，在 Mysql 数据库中自行创建数据表。
3、 运行爬虫代码。
4、 查询 Mysql 数据库的爬取结果数据表。

任务三：数据清洗与分析（30 分）

本阶段的任务：任务二数据采集阶段中完成的酒店网站数据集，其中包含来自不同城市中多家酒店的销售信息，你的小组通过编写代码或脚本完成对文件中酒店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 等开发语言，完成本阶段数据清洗、存储、转化、分析及数据推送等任务。通过多个维度分析酒店的销售信息，并以此评价酒店销售业绩、区域的游客接纳能力、接纳质量等指标。

爬取后的数据文件路径：/h3cu/mysql.csv

3.1 数据清洗

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。 请分析数据集，根据题目规定要求实现数据清洗。

步骤一、酒店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删失或截断。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，从总体上来说分为删除存在缺失值的个案和缺失值插补。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 个的数据条目剔除原始数据集,并输出剔除的条目数量，截图并保存结果。

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//master:9000/hotelsparktask1：

- 解析该文件
- 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
- 程序打包并在 hadoop 平台运行，结果输出至 hdfs 文件系统中//master:9000/hotelsparktask1

根据步骤一要求，完成以下内容：

1) 运行代码，删除数据源中缺失值大于 3 个字段的数据，打印输出删除条目数。
2) 查看清洗后输出的结果文件总行数 /master:9000/hotelsparktask1)。

步骤二、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。

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//master:9000/hotelsparktask2：

- 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
- 程序打包并在 hadoop 平台运行，结果输出至 hdfs 文件系统中//master:9000/hotelsparktask2

根据步骤二要求，完成以下内容：

1) 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。
2) 查看清洗后输出的结果文件(master:9000/hotelsparktask2)总行数。

3.2 数据分析

步骤一、城市游客接纳能力是城市规划建设中的重要指标，其中城市的酒店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酒店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，截图并保存结果。

请根据数据清洗的输出数据集，编写 Mapreduce 程序统计各城市的酒店数量和房间数量，以城市房间数量降序排列并输出前 10 条统计结果，同时创建并写入数据表 table3_1。要求输出字段包含：省份、城市、酒店数量、房间数量。

数据定义如下：

数据项	字段名	备注
省份	province	-
城市	city	-
酒店数量	hotel_num	-
房间数量	room_num	-

数据样式如下：

province	city	hotel_num	room_num
贵州	贵阳	1234	123456.0

根据步骤一要求，完成以下内容：

1） 运行代码，统计各城市的酒店数量和房间数量，以城市房间数量降序排列，并打印输出前 10 条统计结果。

2） 创建表 table3_1。

3） 将统计结果写入表 table3_1 中。

4） 查看表 table3_1 前 5 行数据。

步骤二、OTA，全称为 Online Travel Agency，中文译为“在线旅行社”，是旅游电子商务行业的专业词语。指“旅游消费者通过网络向旅游服务提供商预定旅游产品或服务，并通过网上支付或者线下付费，即各酒店通过网络进行产品营销或产品销售”。OTA 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 OTA 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 OTA 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 OTA 规则擅自以低价让客户取消订单，这种情况又叫做“切单”。OTA 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，OTA 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，截图并保存结果。

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_2。

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

根据步骤二要求，完成以下内容：

1） 创建表 table3_2。

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_2 中。

任务四、数据可视化（20 分）

请根据任务三数据分析的结果，使用 flask 框架，结合 echarts 完成下列题目。可视化文件路径： /h3cu/mysql.csv

1、 出租率是反映酒店经营状况的一项重要指标，它是已出租的客房数与酒店可以提供租用的房间总数的百分比。酒店出租率的情况可以在一定程度上反应出该酒店的整体运营的情况，为了更好的分析指定酒店的入住情况，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。

1） 请编写代码，提取出租率前 10 的城市，并降序排列。

2） 主标题为城市出租率（字体颜色：红色，加粗），副标题为出租率前十的城市（字体颜色：黑色），纵坐标为出租率，横坐标为城市名称（字体 颜色：黑色）。

3） 输出柱状图。

2、 连锁酒店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符合大众化消费。连锁酒店无论在装修、服务还是信誉上都有较大的竞争优势，所以连锁酒店是出差、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，一些已经成熟酒店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，统计指定连锁酒店的经营状况，并以指定图例进行呈现。

请根据以大区划分，统计各地 7 天酒店的出租率（保留 6 位小数），并以折线图呈现。

我国划分大区共有六个：为东北、华北、华东、中南、西北、西南，大区中的省份分布参照下表：

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
台港澳地区	台湾、香港、澳门

根据要求完成以下内容：

1） 请编写代码，提取各地 7 天酒店的出租率。

2） 主标题为全国各地酒店的出租率（字体颜色：红色，加粗），副标题为 7 天酒店的出租率（自定义划分地区）（字体颜色：黑色），纵坐标为出租率，横坐标为地区名称（字体颜色：黑色）

3） 输出折线图。

3、 酒店的间夜量也叫间夜数，是酒店在某个时间段内，房间出租率的计算单位，关于酒店间夜量的计算公式为间夜量=入住房间数*入住天数。例如某酒店今天入住的房间数为 500，则今天的间夜量=500*1=500，而又比如某酒店这个月（30 天）的平均每天入住房间数为 400，则这个月的间夜量=500*1*30=15000。请根据指定表中数据统计酒店间夜数相关数据，并以指定图例进行呈现。

根据要求完成以下内容：

1） 请编写代码，提取酒店间夜数相关数据。

2） 主标题为各城市间夜数（字体颜色：红色，加粗）。

3） 输出各城市间夜数地图散点热力图。

任务五：综合分析（15 分）

1、 从酒店分布维度，对酒店运营情况进行分析，以 7 天酒店为例，分析不同地区的酒店出租率情况；

2、 以北京、上海、四川、广东、海南为例，多维度分析说明几个省份酒店的综合运营情况；分析维度：平均评分、直销拒单率和城市出租率；

3、 从 OTA 平台订单来源角度，以任务四中指定的北京酒店为例，对直销和分销订单的情况进行分析，分析维度：直销订单、分销订单及分销比率；

4、 对 OTA 平台未来拓展合作酒店的方向提出建议；

5、 佐证材料需包括文字描述和图例；