**一、hadoop环境搭建**

要求：所有的运行结果都以截图的方式提交到对应报告中

任务1 adoop HA部署管理

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、将JDK包解压到/usr/local/src路径下，命令复制并粘贴至对应报告中；

2、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并粘贴至对应报告中；

3、从master复制JDK环境变量文件到slave1、slave2节点，命令和结果复制并粘贴至对应报告中；

4、配置SSH密钥登录，实现从master登录到slave1，命令和结果复制并粘贴至对应报告中；

5、Zookeeper配置完毕后，在slave2节点启动Zookeeper，查看Zookeeper运行状态，将命令和结果复制并粘贴至对应报告中；

6、Zookeeper、Hadoop HA配置完毕后，在master节点启动Hadoop，并查看服务进程状态，并将结果复制并粘贴至对应报告中；

7、Hadoop HA配置完毕后，在slave1节点查看服务进程，将命令及结果复制并粘贴至对应报告中。

任务2 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。

任务3 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、将指定路径下的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、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；

3、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；

4、配置hive-site.xml文件，将MySQL数据库作为Hive元数据库。将配置文件相关内容复制并粘贴至对应报告中；

5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命令执行初始化，将初始化结果复制粘贴至对应报告中；

6启动Hive，结果输出复制粘贴至对应报告中。

任务4 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。

**二、数据采集（分别使用scrapy/requests开发）**

要求：

1. 所有爬取的数据保存到数据库中，数据库的表需要自己定义。
2. 所有爬取的数据保存到csv文件中。

任务1 网站地址：https://www.autohome.com.cn/nanjing/cheshi/#pvareaid=3311333

分别爬取热门、SUV、小型车、紧凑型车、中型车、大型车、MPV等6种类型的汽车信息，爬取字段如下所示：



要求：

1. 每一种类型的汽车数据需要单独在数据库中建表，每一种类型的汽车数据需要保存到不同的csv文件中，表名和文件名自拟。
2. 提交每一种汽车类型数据的前10条信息（数据库的截图和csv的截图各一张）
3. 只爬取当前页的数据。

任务2网站地址：https://www.autohome.com.cn/nanjing/cheshi/#pvareaid=3311333

爬取8万以下、8-10万、10-15万等3种类型的汽车信息，爬取字段如下所示：



要求：

1. 每一种类型的汽车数据需要单独在数据库中建表，每一种类型的汽车数据需要保存到不同的csv文件中，表名和文件名自拟。
2. 提交每一种汽车类型数据的前10条信息（数据库的截图和csv的截图各一张）
3. 只爬取当前页的数据。

任务3 网站地址：<https://www.autohome.com.cn/dealer/t3a320100.html?pvareaid=2113326>

爬取字段如下所示：



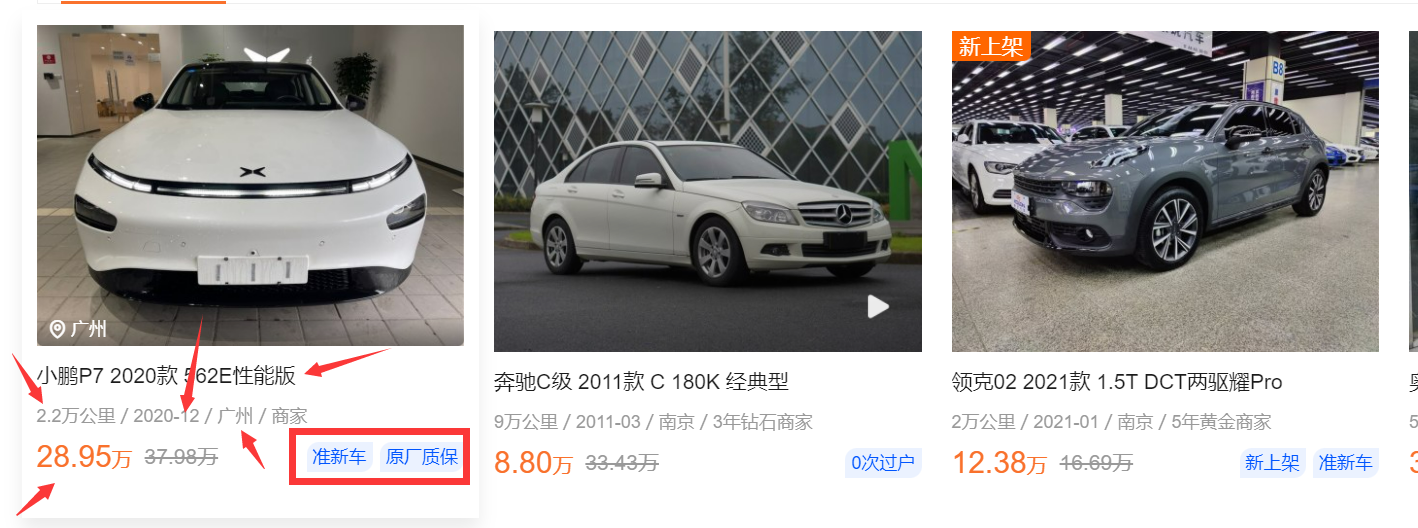
要求：

1）爬取的经销商促销信息不少于100条。

2）提交经销商促销信息的前10条信息、后10条信息、信息总数（数据库的截图和csv的截图各一张）

任务4 网站地址：<https://www.che168.com/nanjing/list/#pvareaid=100533>

爬取字段如下：



要求：

1）爬取的数据不少于100条。

2）提交爬取数据的前10条信息、信息总数（数据库的截图和csv的截图各一张）

**三、数据清洗（使用spark开发，数据放到hdfs上）**

二手房的完整数据为“赛前模拟练习4 二手房data.csv”，二手房的部分数据如下所示：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 小区名字 | 总价 | 户型 | 建筑面积 | 单价 | 朝向 | 楼层 | 装修 | 区域 |
| 0 | 中天北湾新城 | 89万 | 2室2厅1卫 | 89平米 | 10000元/平米 | 南北 | 低层 | 毛坯 | 高新 |
| 1 | 桦林苑 | 99.8万 | 3室2厅1卫 | 143平米 | 6979元/平米 | 南北 | 中层 | 毛坯 | 净月 |
| 2 | 嘉柏湾 | 32万 | 1室1厅1卫 | 43.3平米 | 7390元/平米 | 南 | 高层 | 精装修 | 经开 |
| 3 | 中环12区 | 51.5万 | 2室1厅1卫 | 57平米 | 9035元/平米 | 南北 | 高层 | 精装修 | 南关 |
| 4 | 昊源高格蓝湾 | 210万 | 3室2厅2卫 | 160.8平米 | 13060元/平米 | 南北 | 高层 | 精装修 | 二道 |
| 5 | 金色橄榄城 | 118万 | 3室1厅1卫 | 200平米 | 5900元/平米 | 南北 | 高层 | 简装修 | 二道 |
| 6 | 新星宇和润 | 84万 | 2室1厅1卫 | 75.06平米 | 11191元/平米 | 南北 | 高层 | 精装修 | 南关 |
| 7 | 长电紫盈花城 | 96万 | 2室2厅1卫 | 90平米 | 10667元/平米 | 南北 | 中层 | 精装修 | 二道 |
| 8 | 万科上东区 | 107万 | 3室2厅1卫 | 99.44平米 | 10760元/平米 | 南 | 中层 | 精装修 | 二道 |
| 9 | 中海莱茵东郡 | 205万 | 3室2厅2卫 | 156平米 | 13141元/平米 | 南北 | 高层 | 精装修 | 净月 |
| 10 | 诺睿德国际商务广场 | 105万 | 3室2厅1卫 | 92.7平米 | 11327元/平米 | 南北 | 中层 | 精装修 | 净月 |
| 11 | 华润橡树湾 | 198万 | 4室2厅2卫 | 142.1平米 | 13934元/平米 | 南北 | 中层 | 精装修 | 南关 |
| 12 | 二二八厂宿舍 | 65万 | 2室1厅1卫 | 62.93平米 | 10329元/平米 | 南北 | 高层 | 简装修 | 朝阳 |
| 13 | 平阳小区 | 56万 | 2室1厅1卫 | 70.1平米 | 7989元/平米 | 南北 | 中层 | 精装修 | 南关 |
| 14 | 上东城市之光 | 85万 | 2室1厅1卫 | 89平米 | 9551元/平米 | 南北 | 中层 | 精装修 | 二道 |
| 15 | 中海国际社区 | 285万 | 4室3厅3卫 | 169.2平米 | 16844元/平米 | 东 | 高层 | 暂无 | 净月 |

将赛前模拟练习4 二手房data.csv上传到hdfs的/secondhouse目录下，然后完成如下任务：

任务1 对二手房数据进行清洗，要求如下：

1）将数据中的所有空值删除。

2）将总价、建筑面积、单价字段中的字符删除，只保留数字部分

3）将清洗后的数据放到hdfs的/secondhouse1目录下。

4）程序输出结果如下所示：

原始数据条数：\*\*\*条

空值数据条数：\*\*条

清洗后数据条数：\*\*条

将以上程序的运行截图贴至报告中

5）显示hdfs的/secondhouse下原始数据的前10条记录，将截图贴至报告中。

6）显示hdfs的/secondhouse1下清洗后数据的前20条记录，将截图贴至报告中。

任务2 在任务1 的基础上，计算每个小区二手房的均价（均价=总价/建筑面积），均价保留小数点1位。

程序运行的结果如下：

小区个数：\*\*\*

小区名称：\*\*，均价：\*\*元/米

小区名称：\*\*，均价：\*\*元/米

......

将以上程序的运行截图贴至报告中 (柱状图)

任务3 在任务1的基础上，计算每个户型的数量和建筑面积的平均值

程序运行的结果如下(按户型数量的降序显示户型信息)：

户型：\*\*\*，数量：\*\*个，平均建筑面积：\*\*\*平方米

户型：\*\*\*，数量：\*\*个，平均建筑面积：\*\*\*平方米

......

将以上程序的运行截图贴至报告中

任务4在任务1的基础上，对建筑面积进行排序，找出建筑面积最大的前10个房屋信息

要求：

1）按建筑面积的升序排序

2）将排序后的数据保存到hdfs/area\_house目录下，需要保存的字段包括：小区名称、总价、户型、建筑面积、朝向、楼层、装修、区域

3）显示hdfs/area\_house目录的数据，将截图贴至报告中。

任务5在任务1的基础上，对房屋单价进行排序，找出单价最低的前20个房屋信息

要求：

1）按房屋单价的降序排序

2）将排序后的数据保存到hdfs/price\_house目录下，需要保存的字段包括：小区名称、总价、户型、建筑面积、单价。

3）显示hdfs/price\_house目录的数据，将截图贴至报告中。

任务6 在任务1的基础上，找出满足以下条件的所有房源信息：建筑面积在100-120平方米、房屋装修程度为精装修的、楼层在中低层的。

要求：

1）对满足以上条件的房源按区域进行分组统计，计算每个区域的房源数量，程序运行的结果如下：

区域：\*\*\*，房源数量：\*\*个

区域：\*\*\*，房源数量：\*\*个

......

将程序的截图贴至报告中

2）将满足以上条件的房源导出到hdfs/search\_house目录下，需要保存的字段包括区域、小区名称、建筑面积、装修、楼层、户型。

3）显示hdfs/search\_house目录的前15条数据，将截图贴至报告中。

任务7 房屋的总价、单价和建筑面积的录入可能存在失误，导致数据录入错误，现对房屋的总价、单价和建筑面积的有效性进行核算，核算的公式为：总价=单价\*建筑面积，统计不符合要求的房屋数量，程序运行的结果如下：

无效房屋数量：\*\*\*个

有效房屋数量：\*\*\*个

房屋总量：\*\*\*个。

将程序的截图贴至报告中

注意：以下房屋的数据是正确的：

房屋建筑面积为143平方米、房屋单价为6979元/平米，房屋总价为99.8万元，因为143\*6979=997997元，997997转为万元，再按四舍五入保留1为小数点后的值为99.8。

**四、数据可视化（flask + echarts + mysql）**

在完成数据清洗任务1 的基础上，将hdfs的/secondhouse1目录下的数据导入到mysql数据库中，然后将完成如下任务：

任务1 计算每个小区二手房的均价（均价=总价/建筑面积），均价保留小数点1位。绘制柱状图，柱状图的横坐标为小区名称，纵坐标为二手房的均价。将网页的截图贴至报告中（截图需包含浏览器的完整界面）

任务2 计算每一种户型的房屋数量和每一种户型的均价（均价=总价/建筑面积），均价保留小数点1位。按房屋数量进行排序，房屋数量最高的前5个户型为热门户型。绘制条形图，横坐标为户型的均价，纵坐标为户型的名称。

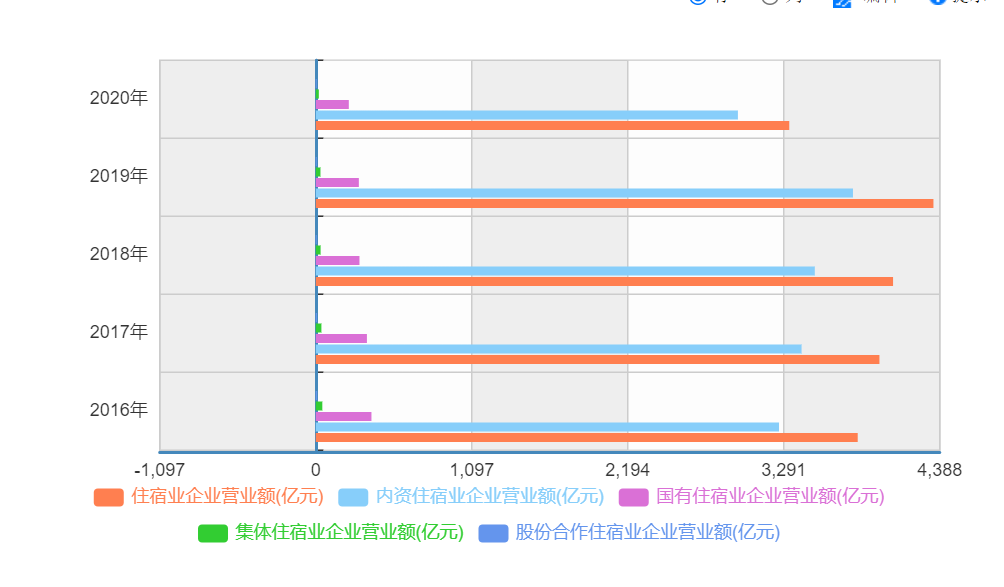
任务3 计算每一个小区的精装修、简装修、毛坯房屋的数量。绘制柱状图，柱状图的横坐标为小区名称，纵坐标为指定小区的精装修、简装修、毛坯房屋的数量。

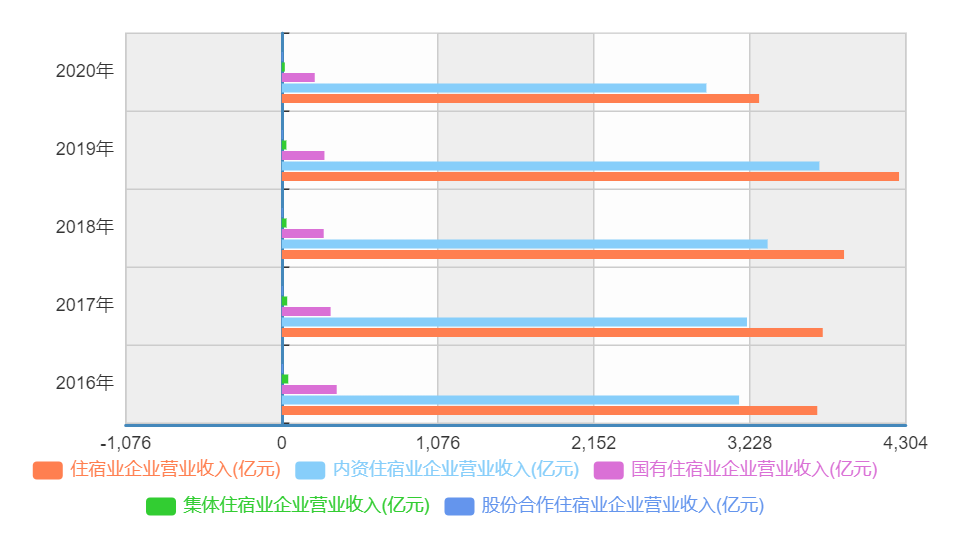
任务4 计算每一个区域的房屋数。绘制饼状图，显示每一个区域的房屋数在总房屋数的占比。

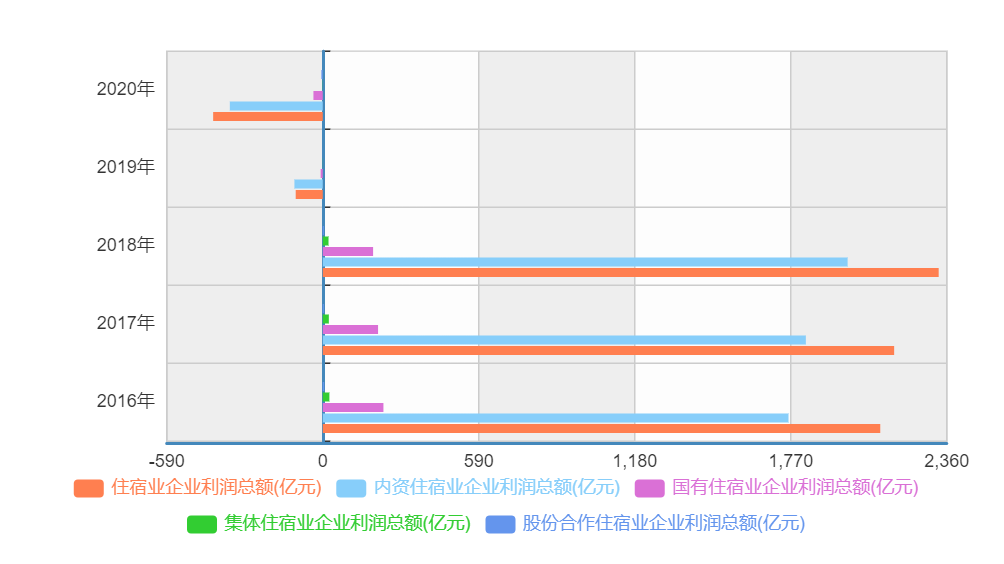
任务5 计算每一个区域的中层、高层、低层的房屋数量。绘制柱状堆叠图，横坐标为区域名称，纵坐标为房屋总值，其中柱状图的下方为低层房屋数量（绿色）、中部为中层房屋数量（红色）、上部为高层房屋数量（黄色）。

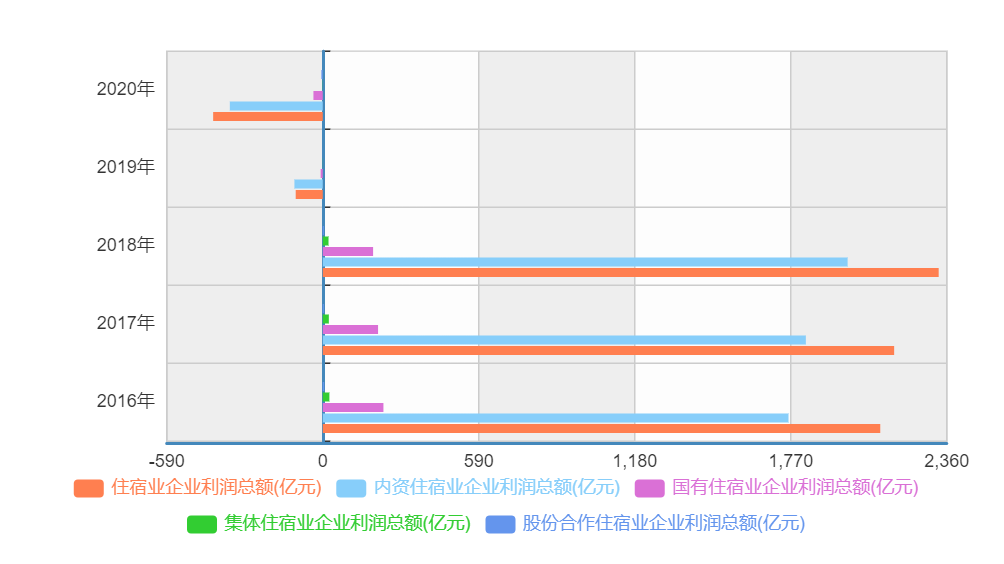
**五、文档分析**

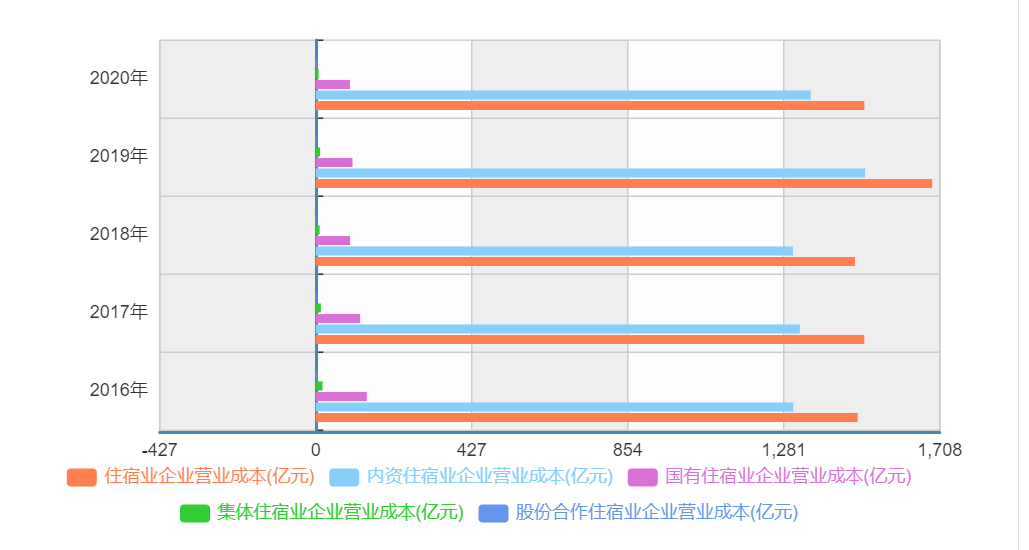
任务1 根据下图分析我国住宿业的基本特征











任务2 根据下图分析我国旅游业的基本特征

