**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 Sqoop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. 解压指定路径下的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,并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sqoop-env.sh文件修改内容复制粘贴至对应报告中。

测试Sqoop连接MySQL数据库是否成功，将执行命令和执行结果复制粘贴至对应报告中。

**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://nanjing.newhouse.fang.com/house/s/

爬取楼盘列表信息如下：



进入到指定楼盘页面，爬取楼盘详情信息如下：



要求：

1）至少爬取2页楼盘列表信息。

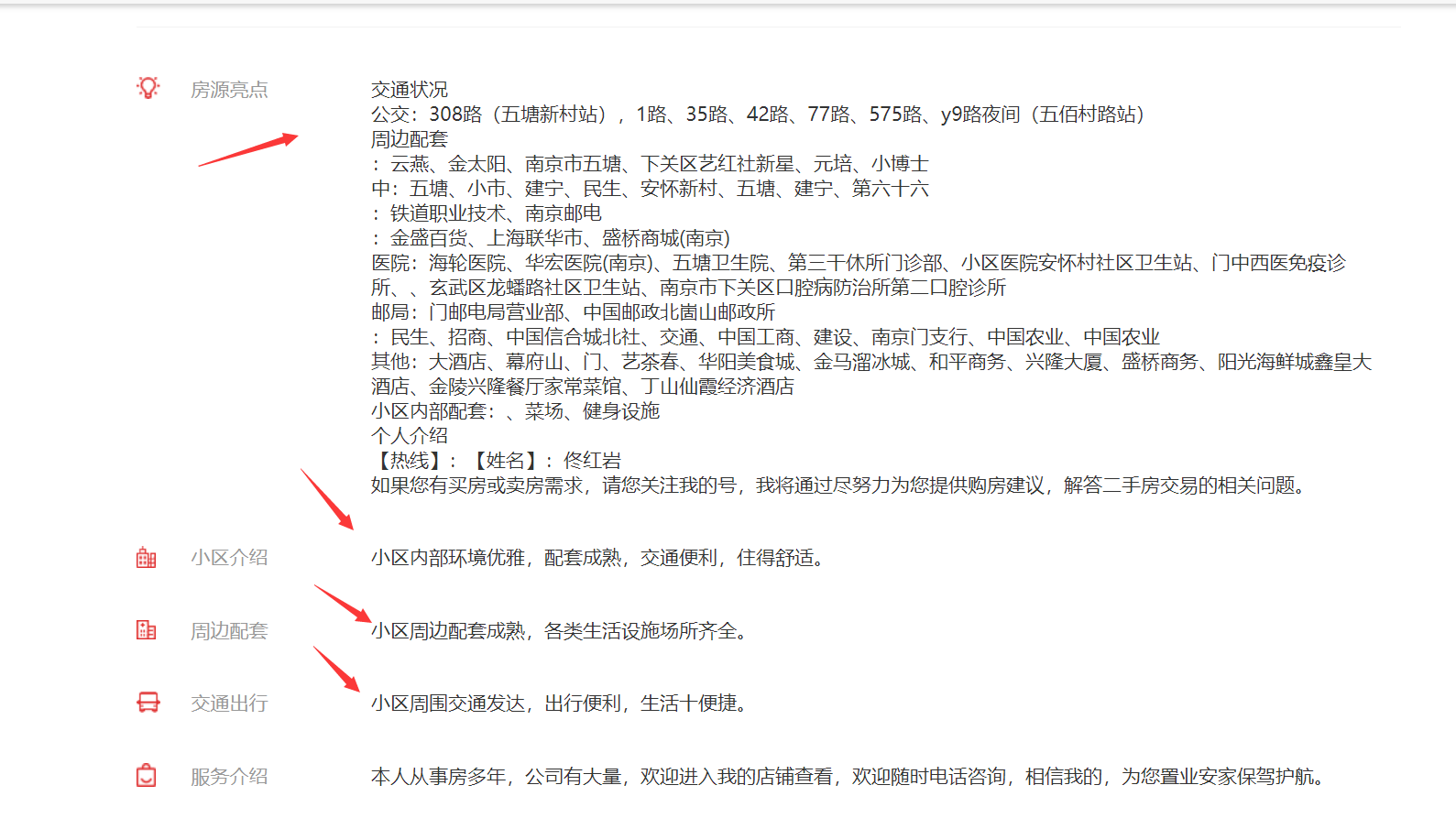
1. 分别提交楼盘列表信息和楼盘详情信息前10条数据（数据库的截图和csv的截图）

任务2网站地址ip：https://nanjing.zu.fang.com/

爬取租房列表信息：



进入到租房详情页面，爬取租房详情信息：



1）至少爬取2页租房列表信息。

1. 分别提交租房列表信息和租房详情信息前10条数据（数据库的截图和csv的截图）

任务3 网站地址：<https://nanjing.esf.fang.com/>

爬取二手房列表信息：



进入二手房详情页面，爬取二手房详情：



1. 至少爬取2页二手房列表信息。

2）分别提交二手房列表信息和二手房详情信息前10条数据（数据库的截图和csv的截图）

**三、数据清洗（使用spark开发，数据放到hdfs上）**

某停车场现有车位100个，2018-1-1至2018-3-31的停车记录保存在“模拟测试5 停车场信息表 清洗题目数据.csv”，该文件的部分数据如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **cn** | **timein** | **timeout** | **price** | **state** | **rps** |
| 赣CFF120 | 2018-01-01 00:03:13 | 2018-01-01 00:23:52 | 3 | 1 | 99 |
| 云N84SU5 | 2018-01-01 00:09:37 | 2018-01-01 00:44:54 | 3 | 1 | 99 |
| 冀RLDH16 | 2018-01-01 00:38:08 | 2018-01-01 00:45:29 | 3 | 1 | 100 |
| 豫K869CW | 2018-01-01 00:52:53 | 2018-01-01 00:59:04 | 3 | 1 | 100 |
| 新QWWA64 | 2018-01-01 01:20:37 | 2018-01-01 01:24:10 | 3 | 1 | 100 |
| 京C511CH | 2018-01-01 01:29:35 | 2018-01-01 01:44:23 | 3 | 1 | 100 |

其中，cn为车牌号，timein为入场时间，timeout为出场时间，price为停车费，state为1表示车辆已缴费驶出，state为0表示车辆还未驶出停车场，rps表示停车场当前剩余车位数。当state为0时，timeout为0，表示该车辆出场时间不确定。

任务1 timein和timeout可能缺失或为无效日期格式，将这些记录删除，有效数据保存到hdfs的/chedata目录下，程序输出结果如下：

原始数据条数：\*\*\*条

无效数据条数：\*\*条

清洗后数据条数：\*\*条

将以上程序的运行截图贴至报告中

任务2 在任务1的基础上，找出所有未驶出停车场的车辆信息。程序输出结果如下：

车牌号：\*\*\* 入场时间：yyyy-mm-dd \*\*:\*\*:\*\* 停车费：

车牌号：\*\*\* 入场时间：yyyy-mm-dd \*\*:\*\*:\*\* 停车费：

......

将以上程序的运行截图贴至报告中

任务3 在任务1的基础上，对所有驶出停车场的车辆计算该车辆在停车场停留的时长，并找出停留时长最长的前5辆车，按降序排列输出这5辆车的信息。程序输出结果如下：

车牌号：\*\*\* 入场时间：yyyy-mm-dd \*\*:\*\*:\*\* 出场时间：yyyy-mm-dd \*\*:\*\*:\*\* 停车时长：\*\*分钟

车牌号：\*\*\* 入场时间：yyyy-mm-dd \*\*:\*\*:\*\* 出场时间：yyyy-mm-dd \*\*:\*\*:\*\* 停车时长：\*\*分钟

......

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，对所有驶出停车场的车辆的停车费进行排序，按升序排列输出前5辆车的信息。程序输出结果如下：

车牌号：\*\*\*入场时间：yyyy-mm-dd \*\*:\*\*:\*\* 停车费：\*\*元

车牌号：\*\*\*入场时间：yyyy-mm-dd \*\*:\*\*:\*\* 停车费：\*\*元

......

将以上程序的运行截图贴至报告中

任务5 在任务1的基础上，计算2018-1-1至2018-1-10期间每一天的停车费总数，并按降序排序输出统计数据。程序输出结果如下：

日期：yyyy-mm-dd 停车费总数：\*\*元

日期：yyyy-mm-dd 停车费总数：\*\*元

.....

将以上程序的运行截图贴至报告中

任务6 在任务1的基础上，计算每一个月的入场停车的总次数，并按升序输出统计数据。程序输出结果如下：

将以上程序的运行截图贴至报告中

月份：yyyy-mm 入场停车总次数：\*\*

月份：yyyy-mm 入场停车总次数：\*\*

......

将以上程序的运行截图贴至报告中

任务7 在任务1的基础上，计算2018-2-1至2018-2-10期间每一天停车场停放车辆最多的是在什么时间。程序输出结果如下：

时间：2018-2-1 \*\*-\*\*-\*\* ，此时停车场停放车辆数量：\*\*

时间：2018-2-2 \*\*-\*\*-\*\* ，此时停车场停放车辆数量：\*\*

......

将以上程序的运行截图贴至报告中

**四、数据可视化（flask + echarts + mysql）**

某停车场现有车位100个，2018-1-1至2018-3-31的停车记录保存在“模拟测试5 停车场信息表 可视化题目数据.csv”，该文件的部分数据如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **cn** | **timein** | **timeout** | **price** | **state** | **rps** |
| 赣CFF120 | 2018-01-01 00:03:13 | 2018-01-01 00:23:52 | 3 | 1 | 99 |
| 云N84SU5 | 2018-01-01 00:09:37 | 2018-01-01 00:44:54 | 3 | 1 | 99 |
| 冀RLDH16 | 2018-01-01 00:38:08 | 2018-01-01 00:45:29 | 3 | 1 | 100 |
| 豫K869CW | 2018-01-01 00:52:53 | 2018-01-01 00:59:04 | 3 | 1 | 100 |
| 新QWWA64 | 2018-01-01 01:20:37 | 2018-01-01 01:24:10 | 3 | 1 | 100 |
| 京C511CH | 2018-01-01 01:29:35 | 2018-01-01 01:44:23 | 3 | 1 | 100 |

其中，cn为车牌号，timein为入场时间，timeout为出场时间，price为停车费，state为1表示车辆已缴费驶出，state为0表示车辆还未驶出停车场，rps表示停车场当前剩余车位数。

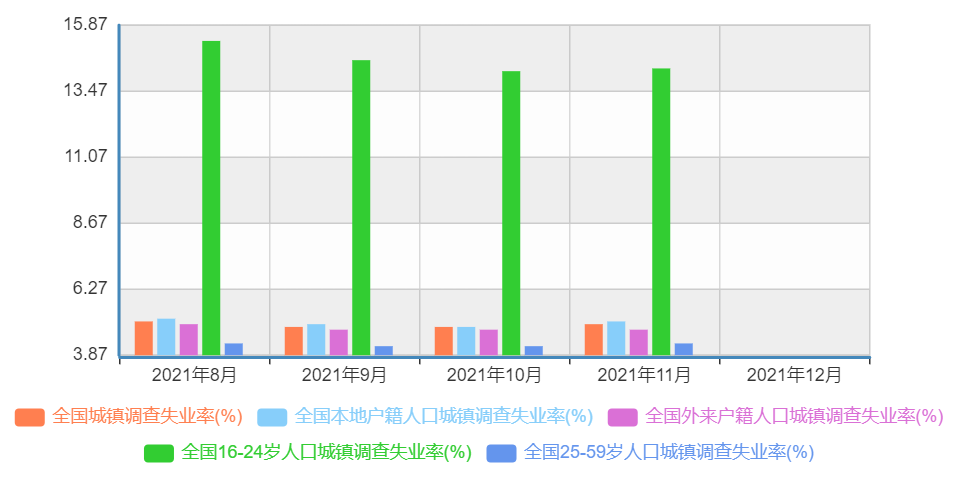
任务1 计算每一个月停车场收取的停车费。绘制柱状图，柱状图的横坐标为月份，纵坐标为停车费。将网页的截图贴至报告中（截图需包含浏览器的完整界面）

任务2计算每一个月的入场停车的总次数。绘制饼状图，显示每一个月的入场停车次数在停车总次数的比例。

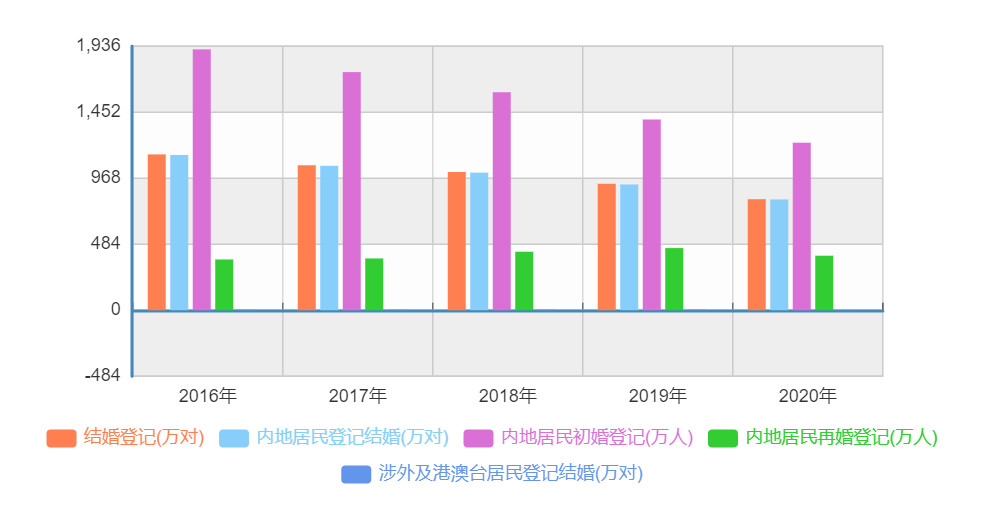
任务3 计算2018-1-1至2018-1-10期间每一天车辆的平均停留时长（时长单位为：分钟。平均停留时长=车辆停车总时长/车辆数）。绘制柱状图，柱状图的横坐标为日期，纵坐标为平均停留时长。将网页的截图贴至报告中（截图需包含浏览器的完整界面）

**五、文档分析**

任务1 请说明我国城镇人口失业率的主要特点



任务2 请说明我国婚姻登记的现状和发展趋势



任务3 请说明我国交通事故的主要特点

