**一、hadoop环境搭建**

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,并将变量内容复制粘贴至对应报告中；

3、修改并配置sqoop-env.sh文件，并将修改内容复制粘贴至对应报告中；

4、测试Sqoop连接MySQL数据库是否成功，结果复制粘贴至对应报告中。

任务5 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。

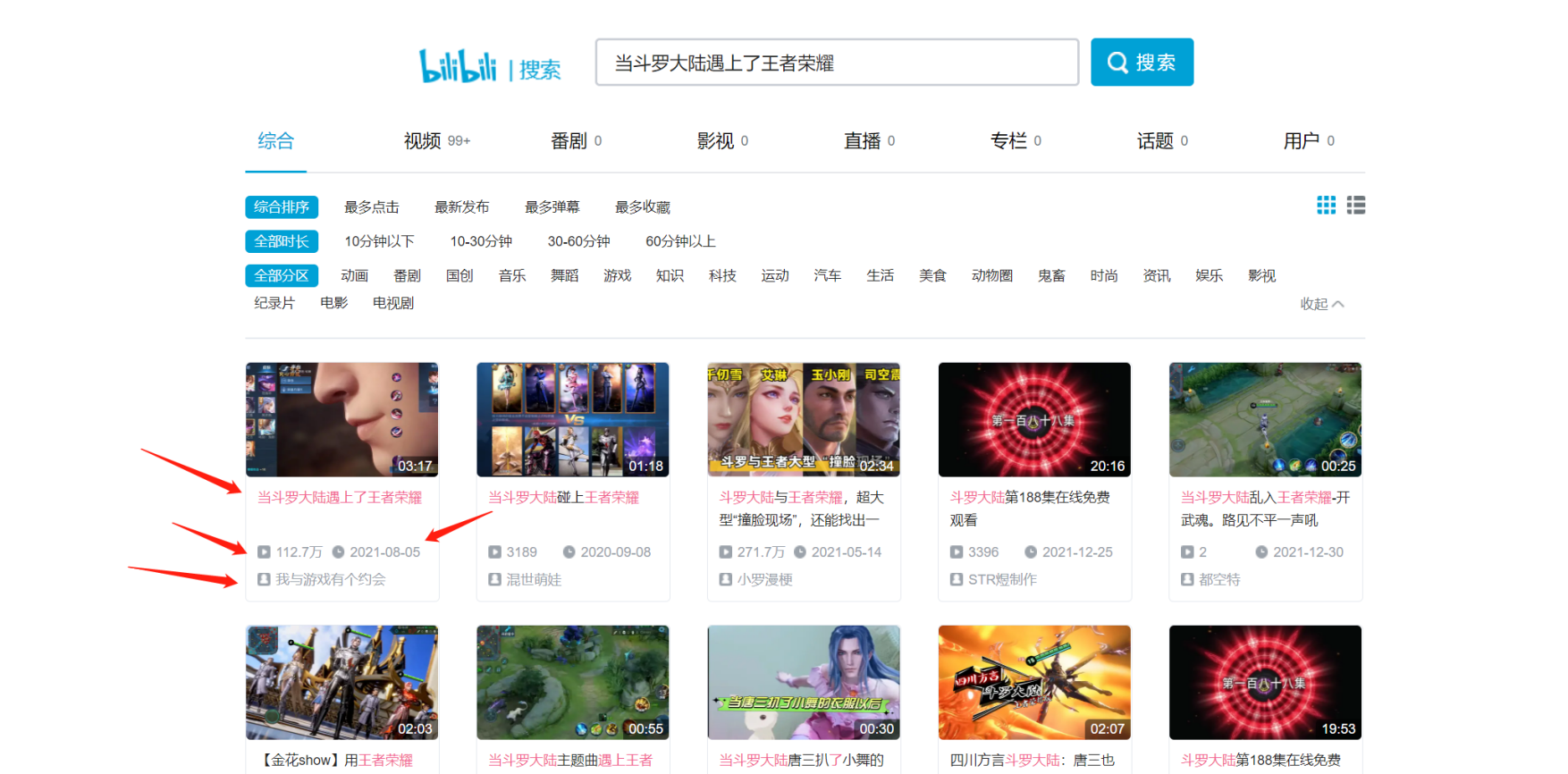
**二、数据采集（分别使用scrapy/requests开发）**

**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json文件中。**

1、爬取B站的信息（至少爬取10页）：<https://search.bilibili.com/all?keyword=%E5%BD%93%E6%96%97%E7%BD%97%E5%A4%A7%E9%99%86%E9%81%87%E4%B8%8A%E4%BA%86%E7%8E%8B%E8%80%85%E8%8D%A3%E8%80%80&from_source=web_search>

需要爬取的数据如方框所示：



进入到每个网页后，需爬取的数据如方框所示：



2、爬取阳光问政数据（至少爬取10页）<https://wz.sun0769.com/political/index/politicsNewest?id=1>

需爬取的数据如方框所示：



点击进入”问政标题”里的标题，进入如下页面，需爬取的数据如箭头和方框所示：



**三、数据清洗（使用spark开发，数据放到hdfs上）**

餐饮外卖数据集waimai.csv的字段包含“餐厅id，城市，客单价，推单数，接单数，整体时长，到店时长，取餐时长，用户投诉数，差评数，”字段，数据集waimai.csv的内容如下：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 餐厅id | 城市 | 客单价 | 推单数 | 接单数 | 整体时长 | 到店时长 | 取餐时长 | 用户投诉数 | 差评数 | 好评数 |
| 111 | 上海 | 12 | 450 | 450 | 10 | 5 | 5 | 4 | 2 | 4 |
| 112 | 上海 | 33 | 345 | 350 | 11 | 11 |  | 5 | 2 | 33 |
| 113 | 上海 |  | 223 | 200 | 8 | 2 | 6 | 3 | 2 | 3 |
| 114 | 上海 | 15 | 224 | 53 | 5 | 3 | 4 | 4 | 0 | 4 |
| 115 | 上海 |  | 33 | 33 | 6 | 3 | 4 | 2 | 3 | 2 |
| 116 | 上海 | 44 | 44 | 44 |  | 3 | 4 | 1 | 1 | 4 |
| 222 | 北京 | 54 | 66 | 66 | 4 | 4 |  | 3 | 13 | 13 |
| 223 | 北京 | 23 | 75 | 4 | 4 | 2 | 2 | 1 | 1 | 1 |
| 224 | 北京 | 44 | 55 | 67 | 5 | 2 | 3 | 1 | 1 | 22 |
| 225 | 北京 | 23 | 422 | 55 | 6 | 4 | 2 | 3 | 1 | 0 |
| 226 | 北京 | 5 | 3 | 66 | 7 | 3 | 5 | 4 | 33 | 3 |
| 227 | 北京 | 32 | 45 | 45 | 6 |  | 3 | 0 | 2 | 44 |
| 228 | 北京 | 5 | 323 | 323 |  | 5 | 6 |  | 2 | 455 |
| 311 | 南京 | 6 | 4 | 4 | 6 |  |  | 3 | 2 | 245 |
| 312 | 南京 | 23 | 5 | 5 | 7 | 3 | 4 | 3 | 1 | 11 |
| 313 | 南京 |  | 44 | 44 |  | 4 |  | 3 | 1 | 23 |
| 314 | 南京 | 45 | 32 | 32 |  |  | 6 | 3 | 4 | 55 |
| 315 | 南京 |  | 23 | 23 | 2 | 1 | 1 | 3 | 7 | 44 |
| 316 | 南京 |  | 55 | 55 | 4 | 2 | 2 | 4 | 2 | 4 |
| 317 | 南京 | 55 | 44 | 44 | 8 | 4 | 4 |  | 5 | 7 |

1、“推单数”是指外卖平台通过顾客点单向商家推送的订单数量，“接单数”为商家根据自身情况，最终选择接受订单的数量。

剔除属性列“推单数”小于“接单数”的异常数据条目，并在程序中以打印语句输出异常条数。将打印输出结果复制并保存至对应报告中。

示例格式：

===数据集初始样本条数为\*\*\*条===

===“推单数”小于“接单数”的异常数据条数为\*\*\*条===

2、客单价是指客户在该商铺下一单的平均支付价格，针对数据集“客单价”属性，统计缺失值数量，并打印输出，将打印输出结果复制并粘贴至对应报告中。

示例格式：===“客单价”属性缺失记录为\*\*\*条，缺失比例\*\*%===

3、当客单价缺失比例小于5%时，对含缺失值数据记录进行删除，同时在对应答案报告中粘贴如下内容：

示例格式：===“客单价”缺失记录已删除===

当缺失比例大于5%时，利用“客单价”属性中位数对缺失值进行填充，并将中位数打印输出，将打印输出结果复制并粘贴至对应报告中：

示例格式：===“客单价”属性中位数为\*\*\*天===

针对相关数据集抽取北京地区相关数据记录，计算商家好评比，将“好评比”作为新属性添加至数据集的最后一列。并查看输出文件前3条记录，将查看命令与执行结果复制粘贴至对应报告中。【好评比计算公式：好评比=好评数/评价数】

4、整体时长=到店时长+取餐时长，对这三列数据进行校验，将异常记录删除，异常数据的判断标准为：1）到店时长字段不为空、取餐时长字段为空的。2）到店时长字段为空、取餐时长字段不为空的。3）到店时长字段为空、取餐时长字段为空，整体时长不为空的。4）到店时长不为空、取餐时长不为空、整体时长为空的。5）到店时长不为空、取餐时长不为空、整体时长不等于到店时长+取餐时长的。

删除符合上述条件的异常记录，打印输出结果复制并粘贴至对应报告中：

示例格式：===“时长异常数据为\*\*\*条===

5、计算每个城市的好评数和差评数，打印出好评数最高的前3个城市的名称和差评师最高的前3个城市名称。

6、找出被用户投诉的所有餐厅id，用户投诉数可能为无效数据，如数据缺失、数值为负值。剔除掉所有无效的用户投诉数，找出所有被用户投诉的餐厅id，并按用户投诉数降序排列，数据保存到hdfs的/tousu文件中。（用户投诉数为0是有效值，表示该餐厅未被投诉）。

7、以客单价降序排列，输出前5条记录到hdfs的/tuidan文件中，每一条记录的字段包括餐厅id、城市、客单价、推单数、接单数。

四、数据可视化（flask + echarts + mysql）

在mysql数据库中创建餐饮外卖表table1，该表的字段包含“餐厅id，城市，客单价，推单数，接单数，整体时长，到店时长，取餐时长，用户投诉数，差评数，好评数”字段，table1的内容如下：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 餐厅id | 城市 | 客单价 | 推单数 | 接单数 | 整体时长 | 到店时长 | 取餐时长 | 用户投诉数 | 差评数 | 好评数 |
| 111 | 上海 | 12 | 450 | 450 | 10 | 5 | 5 | 4 | 2 | 4 |
| 112 | 上海 | 33 | 300 | 300 | 11 | 5 | 6 | 5 | 2 | 33 |
| 113 | 上海 | 32 | 223 | 200 | 8 | 2 | 6 | 3 | 2 | 3 |
| 114 | 上海 | 15 | 224 | 53 | 5 | 3 | 4 | 4 | 0 | 4 |
| 115 | 上海 | 43 | 33 | 33 | 6 | 3 | 4 | 2 | 3 | 2 |
| 116 | 上海 | 44 | 44 | 44 |  | 3 | 4 | 1 | 1 | 4 |
| 222 | 北京 | 54 | 66 | 66 | 4 | 4 |  | 3 | 13 | 13 |
| 223 | 北京 | 23 | 75 | 4 | 4 | 2 | 2 | 1 | 1 | 1 |
| 224 | 北京 | 44 | 55 | 67 | 5 | 2 | 3 | 1 | 1 | 22 |
| 225 | 北京 | 23 | 422 | 55 | 6 | 4 | 2 | 3 | 1 | 0 |
| 226 | 北京 | 5 | 66 | 66 | 7 | 3 | 5 | 4 | 33 | 3 |
| 227 | 北京 | 32 | 45 | 45 | 6 | 3 | 3 | 0 | 2 | 44 |
| 228 | 北京 | 5 | 323 | 323 | 11 | 5 | 6 | 0 | 2 | 455 |
| 311 | 南京 | 6 | 4 | 4 | 6 | 2 | 4 | 3 | 2 | 245 |
| 312 | 南京 | 23 | 5 | 5 | 7 | 3 | 4 | 3 | 1 | 11 |
| 313 | 南京 | 54 | 44 | 44 | 6 | 4 | 2 | 3 | 1 | 23 |
| 314 | 南京 | 45 | 32 | 32 | 7 | 1 | 6 | 3 | 4 | 55 |
| 315 | 南京 | 21 | 23 | 23 | 2 | 1 | 1 | 3 | 7 | 44 |
| 316 | 南京 | 34 | 55 | 55 | 4 | 2 | 2 | 4 | 2 | 4 |
| 317 | 南京 | 55 | 44 | 44 | 8 | 4 | 4 | 2 | 5 | 7 |

Table2记录了每个城市的经纬度值

|  |  |  |
| --- | --- | --- |
| 城市 | 经度 | 维度 |
| 上海 | 1 | 4 |
| 北京 | 3 | 16 |
| 南京 | 8 | 20 |

1、统计每个城市的商家数量，在控制台按照“餐厅数量”降序排列，打印输出城市名称及包含的餐厅数量。

示例格式：

==1: 城市 \*\*\*\*=餐厅数为\*\*\*个===

==2: 城市 \*\*\*\*=餐厅数为\*\*\*个==

2、气泡图的绘制。要求气泡图标题为“餐厅聚集地理位置展示”，横坐标为经度，纵坐标为维度，以该地理位置的商家数量作为气泡半径，绘制气泡图。

3、柱状图的绘制：统计每个城市对应的所有餐厅的用户投诉数、差评数和好评数的均值，并用柱状图呈现，柱状图的横坐标为城市名称、纵坐标为上述三个均值。

4、条形图的绘制：统计每个城市对应的所有餐厅的客单价的均值，并用条形图呈现，条形图的横坐标为客单价的均值，纵坐标为城市名称。

5、饼图的绘制：统计每个城市对应的所有餐厅的推单数的总值，并用饼图呈现。

6、柱状堆叠图的绘制：统计每个城市对应的所有餐厅的差评数和好评数的总值，并用柱状堆叠图呈现，柱状堆叠图的横坐标为城市名称、柱状堆叠图的纵坐标为评论总值，其中柱状图的下方为好评数（绿色）上方为差评数（红色）。

五、文档分析

1、我国货物进出口统计数据如下图所示，请分析下图的含义，并对我国未来的进出口趋势做出判断。



2、中国服务贸易分类月度统计表如下图所示，请分析该图的含义，并说明我国服务贸易的主要特征。



相关背景知识：

1）同比是指本年度（季度，月份）于上一年年度（季度月度）的对比。环比是指本季度或者月度与上一个季度或者月度的对比。

2）服务贸易是一国的法人或自然人在其境内或进入他国境内向外国的法人或自然人提供服务的贸易行为。主要方式有：从一成员境内向任何其他成员境内提供服务；在一成员境内向任何其他成员的服务消费者提供服务；一成员的服务提供者在任何其他成员境内以商业存在提供服务；一成员的服务提供者在任何其他成员境内以自然人的存在提供服务。其中，服务包括商业服务，通信服务，建筑及有关工程服务，销售服务，教育服务，环境服务，金融服务，健康与社会服务。服务贸易大体上有如下分类：

* 劳动密集型服务：旅游、运输等。
* [资本密集型](https://baike.baidu.com/item/%E8%B5%84%E6%9C%AC%E5%AF%86%E9%9B%86%E5%9E%8B" \t "_blank)服务：航空、通讯、建筑等。
* 技术知识密集型服务：金融、计算机和信息服务等。