**背景描述**

随着中国经济的迅速发展，广大人民群众的收入也在不断提高，但是仍有一部分人群需要在经济上获得帮助，每月领取城市居民最低生活保障，为了更好的实现精准扶贫，你所在的小组需要通过给出的“某市低保人群信息表”，通过采集其中低保人口收入，低保人口中失业人数，视力或听力残疾人数等信息，分析各地区低保人群数量，人均收入情况，近几年收入总体情况等信息，为实现精准扶贫给予数据上的佐证。

为完成该项任务，你所在的小组计划选用在业界广泛应用的“Python”语言，作为整个项目的基础语言，并综合利用MySQL、Matplotlib、pandas、Hive等技术提高开发效率并实现项目要求，本次项目环境搭建采用服务器集群方式，通过对“某市低保人群信息表”的采集、清洗与分析，力求实现对精准扶贫给予数据支撑。

# 任务一：Hadoop平台及组件的部署管理（15分）

**1、环境部署**

1）Hadoop系统存储于“/usr/local/adoop” ，要求配置hadoop.tmp.dir目录存放位置为“/usr/local/adoop/tmp”；

2）配置hadoop的dfs.namenode.name.dir为/usr/local/adoop/tmp/dfs/name；

3）配置hadoop的dfs.datanode.data.dir为/usr/local/adoop/tmp/dfs/data；

4）格式化NameNode；

5）开启NameNode和DataNode守护进程。

**2、网络配置（所有节点）**

1）修改当前机器名；

2）退出当前登录，并重新登录；

3）关闭防火墙；

4）修改当前机器IP；

5）配置hosts文件；

6）重启网络；

7）创建一个普通用户（也可以在安装CentOS系统过程中在图形界面创建hadoop用户，密码设置为hadoop）。

**3、SSH无密码验证配置**

Hadoop运行过程中需要管理远端Hadoop守护进程，在Hadoop启动以后，NameNode是通过SSH（Secure Shell）来启动和停止各个DataNode上的各种守护进程的。这就必须在节点之间执行指令的时候是不需要输入密码的形式，因此需要配置SSH运用无密码公钥认证的形式，这样NameNode使用SSH无密码登录并启动DataName进程，同样原理，DataNode上也能使用SSH无密码登录到NameNode。

1）安装和启动SSH协议；（**所有节点**）

2）切换到hadoop用户；

3）每个节点生成秘钥对；（**所有节点**）

4）查看"/home/hadoop/"下是否有".ssh"文件夹，且".ssh"文件下是否有两个刚生产的无密码密钥对；（**所有节点**）

5）把id\_rsa.pub追加到授权的key里面去；（**所有节点**）

6）修改文件"authorized\_keys" 权限；（**所有节点**）

7）设置SSH配置；（**所有节点**）

8）设置完之后记得重启SSH服务，才能使刚才设置有效；（**所有节点**）

9）切换到hadoop用户；

10）验证是否成功；（**所有节点**）

11）把master节点的公钥id\_rsa\_pub复制到每个slave点；

12）在每个slave节点把master节点复制的公钥复制到authorized\_keys文件；（**所有slave节点**）

13）删除id\_rsa.pub文件；（**所有slave节点**）

14）验证master到每个slave节点无密码验证；（**master节点**）

15）每一个slave节点的公钥复制到master；（注意15、16、17步骤完成一个slave节点后再操作下一个）

16）在master节点把从slave节点复制的公钥复制到authorized\_keys文件；（**master节点**）

1. 删除id\_rsa.pub文件；（**master节点**）

18）验证每个slave节点到master无密码验证；（**slave节点**）

**4、Java环境安装（所有节点都要配置）**

将jdk-8u77-linux-x64.tar.gz包上传到master节点/root目录下。

1）切换到root用户；

2）新建java目录；

3）解压到/usr/java目录下；

4）配置环境变量；

5）使添加的环境变量生效；

6）验证安装成功。

**5、在Master节点上安装hadoop**

1）解压缩到/usr目录下；

2）重命名；

3）配置hadoop环境变量；

4）使配置的hadoop的环境变量生效；

5）配置hadoop-env.sh；

6）配置core-site.xml；

7）配置hdfs-site.xml；

8）配置yarn-site.xml；

9）配置mapred-site.xml；

10）配置masters文件；

11）配置slaves文件；

12）新建目录；

13）修改/usr/local/hadoop目录的权限；

14）将master上的hadoop安装文件同步到slave1 slave2；

15）在每个slave节点上配置hadoop的环境变量；（所有slave节点）

16）使配置的hadoop的环境变量生效；（所有slave节点）

17）修改/usr/local/hadoop目录的权限；（所有slave节点）

18）切换到hadoop用户。（所有slave节点）

**6、测试**

1）切换到hadoop；（master节点）

2）先格式化；（master节点）

3）启动hadoop；（master节点）

4）查看Java进程；

5）使用浏览器浏览Master节点机查看NameNode节点状态；

6）浏览Datanode数据节点；

7）使用浏览器浏览Master节点查看所有应用；

8）浏览Nodes；

9）关闭hadoop。

# 任务二、数据采集（15分）

根据下列表头样式，自定义编写或采集可用数据源，并保存到对应“任务二”服务器中。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | Area\_county | Street | Committee | Name1 | Name2 | Id\_number | Identity\_categories | Physical\_condition | Approval\_time | Money |
| 35029 | 铁西区 | 七路街道 | 育工社区 | 徐洪生 | 徐洪生 | 21012119461103xxxx | 无领取离退休金和养老保险金老年人 | 健全 | 2016/07/01 | 30 |
| 35030 | 铁西区 | 七路街道 | 育工社区 | 徐洪生 | 杨凤云 | 21012119701124xxxx | 未登记失业人员 | 健全 | 2016/07/01 | 30 |
| 35031 | 铁西区 | 七路街道 | 育工社区 | 徐洪生 | 王慧萍 | 21010620021123xxxx | 学龄前儿童 | 健全 | 2016/07/01 | 713 |
| 35032 | 铁西区 | 七路街道 | 育工社区 | 徐洪生 | 徐思华 | 21010620011001xxxx | 学生 | 健全 | 2016/07/01 | 407 |
| 35033 | 铁西区 | 七路街道 | 育工社区 | 徐洪森 | 徐文龙 | 21010619960924xxxx | 其他人员 | 一级肢体残疾 | 2016/07/01 | 468 |
| 35034 | 铁西区 | 七路街道 | 育工社区 | 徐洪森 | 徐文静 | 21010619661024xxxx | 登记失业人员 | 健全 | 2016/07/01 | 285 |
| 35035 | 铁西区 | 七路街道 | 育工社区 | 徐洪森 | 徐洪森 | 21010619670804xxxx | 未登记失业人员 | 健全 | 2016/07/01 | 285 |
| 35036 | 铁西区 | 七路街道 | 育工社区 | 徐洪明 | 徐洪明 | 21010620000906xxxx | 学生 | 健全 | 2016/07/01 | 349 |
| 35037 | 铁西区 | 七路街道 | 育工社区 | 徐洪军 | 徐洪军 | 21010619630127xxxx | 未登记失业人员 | 其它 | 2016/07/01 | 227 |
| 35038 | 铁西区 | 七路街道 | 育工社区 | 徐洪军 | 徐祎滢 | 21011419610222xxxx | 未登记失业人员 | 健全 | 2016/07/01 | 457 |
| 35039 | 铁西区 | 七路街道 | 育工社区 | 徐洪军 | 刘晶 | 21010619590729xxxx | 登记失业人员 | 健全 | 2016/07/01 | 457 |
| 35040 | 铁西区 | 七路街道 | 育工社区 | 徐洪国 | 徐洪国 | 21110219500622xxxx | 登记失业人员 | 其它 | 2016/07/01 | 579 |
| 35041 | 铁西区 | 七路街道 | 育工社区 | 徐宏琴 | 徐宏琴 | 21110219490208xxxx | 登记失业人员 | 健全 | 2016/07/01 | 578 |
| 35042 | 铁西区 | 七路街道 | 育工社区 | 徐宏春 | 徐宏春 | 21010619940509xxxx | 其他人员 | 二级精神残疾 | 2016/07/01 | 475 |
| 35043 | 铁西区 | 七路街道 | 育工社区 | 徐宏斌 | 徐宏斌 | 21010619640908xxxx | 未登记失业人员 | 二级视力残疾 | 2016/07/01 | 474 |
| 35044 | 铁西区 | 七路街道 | 育工社区 | 徐宏 | 徐宏 | 21010619670501xxxx | 登记失业人员 | 四级智力残疾 | 2016/07/01 | 474 |

# 任务三、数据清洗与分析（25分）

1. 创建hive表格；
2. 将爬取数据导入相应表；
3. 读取数据集；
4. 清洗数据中的无效数据；
5. 统计指定属性列的缺失值个数；
6. 查看具有缺失值的数据行；
7. 补充Money中缺失值并新保存表；
8. 读入及查看下列数据集。
9. 2013-2015低保人口的收入平均值；
10. 统计2016年各区的低保人口中失业人数；
11. 2015-2016年，“未登记失业人员”的平均收入；
12. 2014年各区具有“视力或听力残疾”的人数；
13. 对指定属性进行标准化，并写入相应文件。

# 任务四、数据可视化（20分）

将分析后的数据推送到MySQL数据库中，并对内容可视化呈现：

1. 使用Matplotlib绘制一个区县的统计图；
2. 展示某两年低保人口的收入平均值；
3. 统计某年各区的低保人口中失业人数；
4. 展现某两年“未登记失业人员”的平均收入；
5. 展现某年各区具有“视力或听力残疾”的人数。

# 任务五：综合分析（15）

根据可视化图表回答以下问题

1. 2016年哪个区域需要的投入的低保经费最多；
2. 2016年哪个区域人口平均收入最低；
3. 2016年哪个区域人口平均收入最高；
4. 如何提高低保人口收入平均值。