练习题三

任务一：Hadoop 平台及组件的部署管理（15 分）

一、 环境部署

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

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

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

4）格式化 NameNode

5）开启 NameNode 和 DataNode 守护进程

二、 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）

三、 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**）

17）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**）

四、 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

五、 在 Master 节点上安装 hadoop

1）解压缩到/usr 目录下

2）重命名

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

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）

六、 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
350229	铁西区	七路街道	育工社区	徐洪生	徐洪生	210121119461103xxxx	无领取离退休金和养老保险金老年人	健全	2016/07/01	30
350330	铁西区	七路街道	育工社区	徐洪生	杨凤云	210121119701124xxxx	未登记失业人员	健全	2016/07/01	30
35	铁西区	七路	育工	徐洪	王慧	2101062 0021123	学龄前儿童	健全	2016/07/01	71

0 3 1		街 道	社 区	生	萍	xxxx				3
3 5 0 3 2	铁西 区	七 路 街 道	育 工 社 区	徐 洪 生	徐 思 华	2101062 0011001 xxxx	学生	健全	2016/ 07/01	4 0 7
3 5 0 3 3	铁西 区	七 路 街 道	育 工 社 区	徐 洪 森	徐 文 龙	2101061 9960924 xxxx	其他人员	一级肢 体残疾	2016/ 07/01	4 6 8
3 5 0 3 4	铁西 区	七 路 街 道	育 工 社 区	徐 洪 森	徐 文 静	2101061 9661024 xxxx	登记失业人员	健全	2016/ 07/01	2 8 5
3 5 0 3 5	铁西 区	七 路 街 道	育 工 社 区	徐 洪 森	徐 洪 森	2101061 9670804 xxxx	未登记失业人 员	健全	2016/ 07/01	2 8 5
3 5 0 3 6	铁西 区	七 路 街 道	育 工 社 区	徐 洪 明	徐 洪 明	2101062 0000906 xxxx	学生	健全	2016/ 07/01	3 4 9
3 5 0 3 3	铁西 区	七 路 街	育 工 社	徐 洪 军	徐 洪 军	2101061 9630127 xxxx	未登记失业人 员	其它	2016/ 07/01	2 2 7

7		道	区							
350038	铁西区	七路街道	育工社区	徐洪军	徐祎滢	210114196102222xxxx	未登记失业人员	健全	2016/07/01	457
350039	铁西区	七路街道	育工社区	徐洪军	刘晶	210106195907290xxxx	登记失业人员	健全	2016/07/01	457
350040	铁西区	七路街道	育工社区	徐洪国	徐洪国	211102195006220xxxx	登记失业人员	其它	2016/07/01	579
350041	铁西区	七路街道	育工社区	徐宏琴	徐宏琴	211102194902080xxxx	登记失业人员	健全	2016/07/01	578
350042	铁西区	七路街道	育工社区	徐宏春	徐宏春	210106199405090xxxx	其他人员	二级精神残疾	2016/07/01	475
350043	铁西区	七路街道	育工社区	徐宏斌	徐宏斌	210106196409080xxxx	未登记失业人员	二级视力残疾	2016/07/01	474

350444	铁西区	七路街道	育工社区	徐宏	徐宏	210106196705010xxx	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（30 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

a) 2013-2015 低保人口的收入平均值

b) 统计 2016 年各区的低保人口中失业人数

c) 2015-2016 年，“未登记失业人员”的平均收入

d) 2014 年各区具有“视力或听力残疾”的人数

e) 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） 哪个区域人口平均收入最低？

2） 哪个区域人口平均收入最高？

3） 如何提高低保人口收入平均值？