课程集群使用说明（2019 年春季)

版本号 ：V1.0

更新时间 ：2019.4.12

撰写人：汪浩港

一、集群使用注意事项

1. 集群仅供课程实验和课程设计使用，不允许利用集群完成其他任务。严禁使用集群从事任何违法行为，严禁存储、发布任何反动，暴力，色情，违法，侵权以及不符合相关政策规定的内容。 助教将对不当使用集群的情况进行包括停用、禁用集群的处理。
2. 集群仅在课程期间有效。在课程设计报告提交截止日期之后，管理员将对集群上的数据进行清理，将删除本地文件系统和 HDFS 上的文件。因此重要的程序和数据请做好备份。
3. 尽量不要往上 Hadoop 集群上往放大量的小碎文件，如果要临时存存储，记得用完以后尽快删除。公用的数据也会在放在 HDFS 的/data 目录下，各个用户不要再重复传数据上去，以节省资源，否则管理员会看情况删除。

4. 由于集群只有一个节点与校园网连接,其他节点都是在集群内部的局域网里,实验时数据一般先是传到主节点的 Linux 文件系统里，然后拷到 HDFS 里的，文件比较大的话，拷好了没意外了就最好尽快删掉，以免浪费空间。
5. 集群使用过程中遇到问题，请在课程 QQ 群询问。

二、集群远程访问

注意：请先在自己的本地机器上熟悉 Linux、Hadoop 的基本操作（尤其是命令行操作），小的练习可以在本地进行，当有大的数据处理任务时再到集群服务器上运行。数据量很小时，在集群上可能不会有明显速度优势，反而可能会因为系统开销而运行更慢。

1. 集群软件版本

- JDK 1.7
- Hadoop 2.7.1

请按上述软件版本编译程序。

2. 登入集群

集群提供 SSH 的远程登录方式。

集群主节点 IP 地址：114.212.190.95 (仅限校园网访问)。用户名、密码将通过其他渠道公布。

通过在本地运行 ssh 命令登录远程集群： ssh [用户名]@114.212.190.95

登录之后将进入集群主节点的远程命令行终端。

登录集群之后，可以随时使用 passwd 命令修改帐号的密码。

3. 与集群交换文件

集群同时提供了 scp 和 sftp 两种传输文件的方式。

- scp 方式：通过 scp 命令将本地文件上传到集群主节点的自己的 HOME 目录下。或者利用 scp 从集群远程下载文件。
- sftp 方式：在 FileZilla 客户端中，主机地址输入 sftp://114.212.190.95，用户名/密码是集群登录帐号和密码，端口号不填。连接成功后，就可以和本地进行文件交互。

通过这两种方式上传/下载的文件都是保存在集群主节点的本地文件系统下。

如果需要将文件从本地文件系统进一步上传/下载到 HDFS，使用 hadoop fs -put/get 等 HDFS 操作命令完成。

4. 数据的存放与用户目录

课程实验所需要的数据会放在公共目录/data 下，注意这个是 **HDFS** 里的目录，不是 Linux 本地文件系统上的目录。

用如下命令查看：

[user-student@master01 ~]$ hadoop fs -ls /data

公用的武侠小说集在

/data/wuxia_novels

每一个用户在 HDFS 里都有一个默认的用户主目录 /user/用户名/ 。当执行命令时若不特别指定，对应的默认目录就是这个目录。例如用户 2019st01 的主目录是/user/2019st01/。

如果运行任务 wordcount

[2019st01@master01 ~] $ hadoop jar wordcount.jar in_dir out_dir

那么对应的目录就是/user/2019st01/in_dir, /user/2019st01/out_dir

要使用公共目录/data 下的数据，要这样写全：

[2019st01@master01 ~]hadoop jar wordcount.jar /data/wuxia_novels out_dir

要注意的是:

- 尽量不要往上 Hadoop 集群上往放大量的小碎文件，如果要临时存存储，记得用完以后尽快删除。
- 公用的数据也会在每个实验放在公共目录/data 下，各个用户不要再重复传数据上去，以节省资源，否则管理员会看情况删除。
- 由于集群只有一个节点与校园网连接,其他节点都是在集群内部的局域网里,实验时数据一般先是传到主节点的 Linux 文件系统里，然后拷到 HDFS 里的，文件比较大的话，拷好了没意外了就最好尽快删掉，以免浪费空间。

5. 集群作业提交

使用 hadoop jar [jar 包路径] [其他参数] 命令提交程序给 Hadoop 执行。

6. 集群 Web UI 访问

集群的 Hadop 已经开启 Web UI 界面，可在此页面上查看任务执行的情况。

访问 Web UI，需要在浏览器中开启集群访问代理：

代理类型：HTTP，代理地址：114.212.190.95，代理端口号：3128，无用户

名和密码

在设置好集群访问代理后，Hadoop Web UI 地址为：http://master01:8088
