**电 子 科 技 大 学**

**实** **验**

**报** **告**

**学生姓名：** 赵则橦

**学号：2017221305013**

**指导教师：** 黄俊

**实验地点：信软楼**

**实验时间：** 2019/11

**一、 实验名称：**Hadoop伪分布式环境模式

**二、** **实验学时：**2学时

**三、 实验目的：**

1. 掌握 Hadoop 的伪分布式环境搭建
2. 熟悉 Hadoop、Linux 的基本命令。

**四、 实验原理**

Hadoop 可以在单节点上以伪分布式的方式运行，Hadoop 进程以分离的 Java 进程来运行，节点既作为 NameNode 也作为 DataNode。也就是说我的设备同时是 NameNode 同时也是 DataNode，在安装配置 Hadoop 的时候，不管是安装本地模式，还是安装分布式模式，我们都需要对其进行配置，其中，core-site.xml、yarn-site.xml、mapred-site.xml、hdfs-site.xml 就是我们需要配置的文件之一。启动 hfds 需要两个命令 start-yarn.sh 和 strat-dfs，同时也可以使用 strat-all.sh 启动所有相关的 hdfs 服务。最后需要格式化 hdfs，格式化 hdfs

的原理有点像我们使用硬盘之前要对硬盘进行格式化，自动进行一些相关的配置。之后在 hdfs 上运行 wordCount，这次运行的 wordCount 与第一次实验就有很大的不同，这次的运行无论输入还是输出都是在 hdfs 上的，不是在本机上。

**五、 实验内容**

首先对 Hadoop 进行相关文件的配置，这是运行 Hadoop 的前提条件，之后

再对 hdfs 进行格式化，这也是运行 hadoop 的前提条件，之后我们就可以进行启动 Hadoop 的步骤，如果前面两个步骤失败的话，这个步骤也不能成功运行最后验证整个实验是否成功的关键一步就能否在 hdfs 创建文件，如果成功那么就可以执行全个实验的最后一步，运行 wordCount 并查看结果。

**六、 实验设备及环境**

Ubuntu

jdk 1.8

hadoop 2.7.X

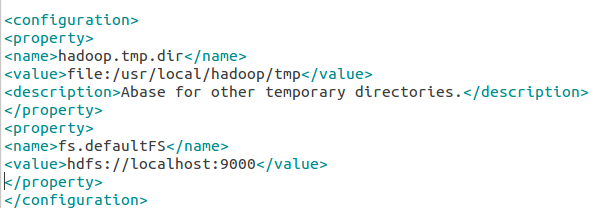
**七、 实验步骤**

1. 配置 core-site.xml。
2. 配置 yarn-site.xml。
3. 创建和配置 mapred-site.xml。
4. 配置 hdfs-site.xml。
5. 格式化 hdfs。
6. 启动 Hadoop.
7. 运行 Hadoop 伪分布式实例

**八、 实验结果与分析（含重要数据结果分析或核心代码流程分析）**

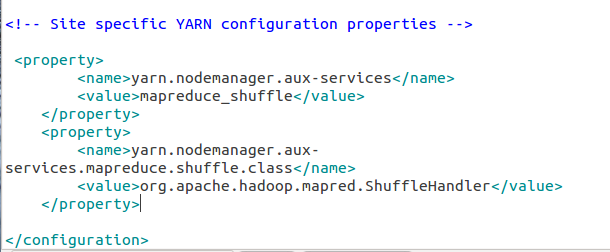
1. 配置 core-site.xml

/usr/local/hadoop/etc/hadoop/core-site.xml 包含了 hadoop 启动时的配置信息。编辑器中打开此文件，并对其进行配置



1. 配置 yarn-site.xml

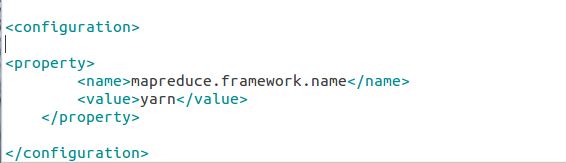
/usr/local/hadoop/etc/hadoop/yarn-site.xml 包含了 MapReduce 启动时的配置信息。编辑器中打开此文件,并进行配置



1. 创建和配置 mapred-site.xml

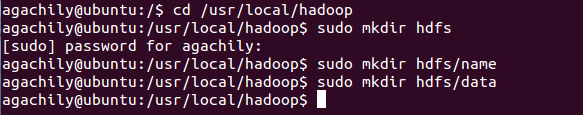
默认情况/usr/local/hadoop/etc/hadoop/文件夹下有 mapred.xml.template 文件，复制该文件，并命名为 mapred.xml，该文件用于指定 MapReduce 使用的框架。

首先复制并重命名，然后编辑器打开此新建文件

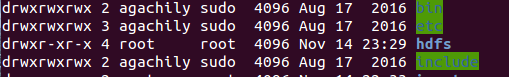


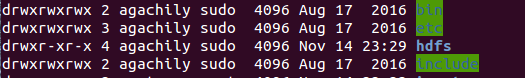
1. 配置 hdfs-site.xml

/usr/local/hadoop/etc/hadoop/hdfs-site.xml 用来配置集群中每台主机都可用，指定主机上作为 namenode 和 datanode 的目录, 创建文件夹

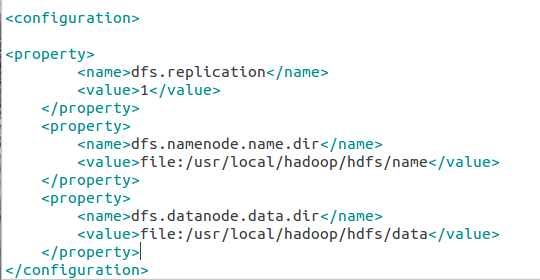


检查文件夹授权是否是，如果不是则修改

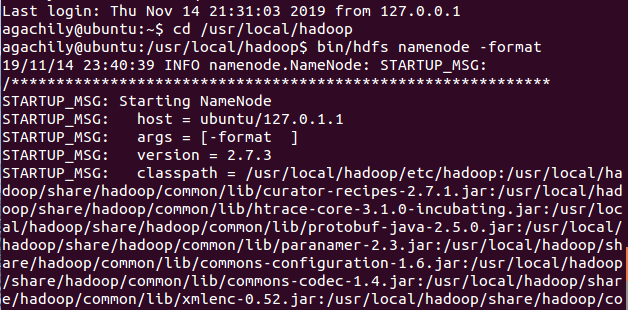




编辑器打开 hdfs-site.xml，进行配置



1. 格式化 hdfs

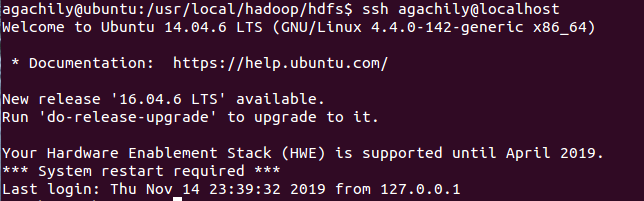


1. 启动hadoop

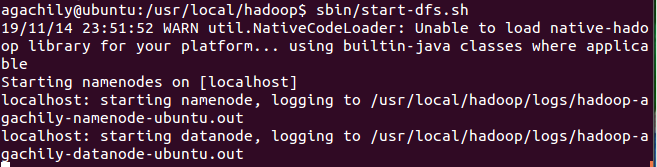
经过上文所描述配置和操作后，一定要检查所有/usr/local/hadoop/下文件的权限，特别是/hdfs/data 和/hdfs/name, 保证为agachily 的权限。下面就可以启动这个单节点的集群



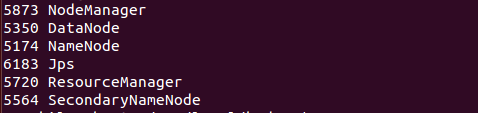
启动这个单节点的集群



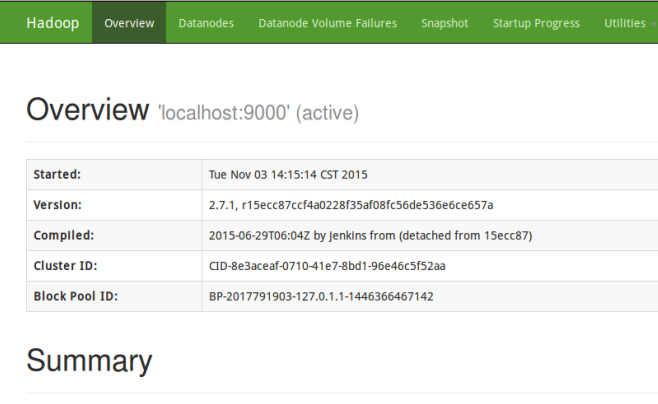
执行启动命令：



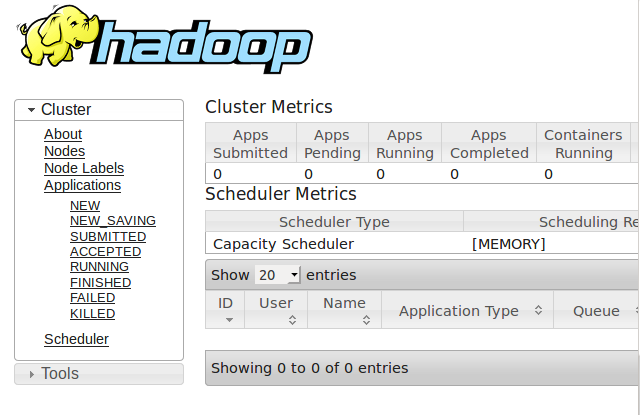
看到Hadoop相关的进程



浏览器打开 <http://localhost:50070/，会看到hdfs管理页面>



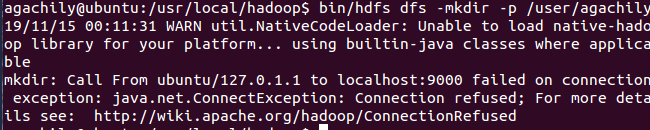
浏览器打开 http://localhost:8088，会看到 hadoop 进程管理页面



1. 运行Hadoop伪分布式实例

伪分布式读取的是 HDFS 上的数据。

要使用 HDFS，首先需要创建用户目录：



接着将 input 中的文件作为输入文件复制到分布式文件系统中，即将 /usr/local/hadoop/input 复制到分布式文件系统中的 /user/yangpan/input 中

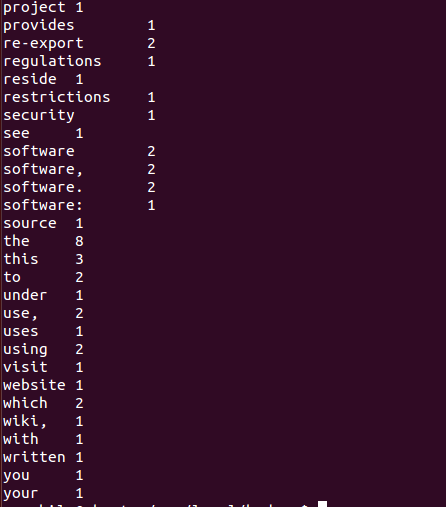
复制完成后，查看文件列表



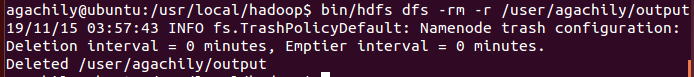
伪分布式运行 MapReduce 作业的方式跟单机模式相同，区别在于伪分布式

读取的是 HDFS 中的文件（可以将单机步骤中创建的本地 input 文件夹删掉以及运行结果来验证这一点）

查看运行结果的命令（查看的是位于 HDFS 中的输出结果）



Hadoop 运行程序时，默认输出目录不能存在，因此运行需要执删除 output 文件夹:



1. **心得体会**

Hadoop可以在单节点上以伪分布式的方式运行。所谓伪分布式就是hadoop进程以分离的Java进程来运行，节点即作为Namenode也作为DataNode。也就是说我们的设备同时是NameNode和DataNode。在进行这次实验的时候，我遇到了很为难的经历。第一次在马上要把实验做完的时候，虚拟机突然卡住重新启动是显示引导程序错误，无法启动，我无奈只好将虚拟机删除再重新安装，并进行环境配置，结果在进行第二次配置时，hadoop安装包解压错误，我配置到后面才发现又出了问题，我只好再次将hadoop卸载。我发现在linux下虽然安装软件比windows麻烦需要进行各种的环境配置，但相对的，在卸载软件时会更简单，只需要将安装包删除即可。

由于对实验指导书中所提到的hdfs不太了解，我于是对这个概念进行了相关的查询，hdfs是一个抽象的文件系统概念，它以流式数据访问模式来存储超大文件，运行于商业硬件集群上。也就是在hadoop中进行处理的文件是存储在hdfs上的。在hadoop集群中主要有两类节点，一类是namenode节点，管理文件系统的命名空间，它同时还维护着文件系统树和整颗树内所有的文件和目录。Datanode则用于存储并检索数据块，并定期向namenode发送所存储的块的列表。

我一开始使用jps的时候并没有查到namenode节点，后来我同过上网查找找到了解决相关问题的办法，链接如下：

<https://blog.csdn.net/hzdxw/article/details/51960211>主要是通过对namenode进行格式化来解决。

对于如何使用hdfs，我也有了相关的了解，掌握向hdfs上上传文件和删除文件的方法。我也了解到伪分布式和单机模式下运行MapReduce的方式相同，但是区别则在于伪分布式读取的是hdfs中的文件。

在本次实验中我成功配置了hadoop的相关配置文件，格式化了hdfs，namenode，成功启动了hdfs并且运行了wordcount程序，掌握了Hadoop的伪分布式环境搭建。

**十、对本实验过程及方法、手段的改进建议：**

我希望能够得到更为清楚明白的讲解和实验指导书。

**报告评分：**

**指导教师签字：**