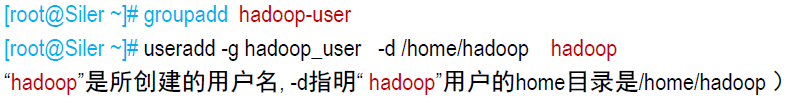
安装单机Hadoop系统与WordCount程序实验

1. 实验环境

Ubuntu 12.04.2 Java 1.8.0\_102 Maven 3.3.9 IntelliJ 15.06

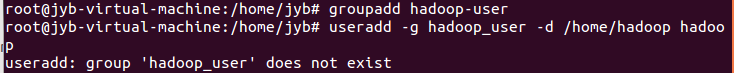
1. 系统安装运行
2. 创建用户组和用户

PPT上为：



但实际按照这样做的话会遇到几个问题

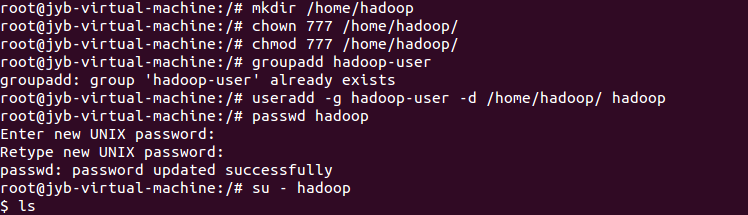
1. 用户组不一致 hadoop-user 和hadoop\_user



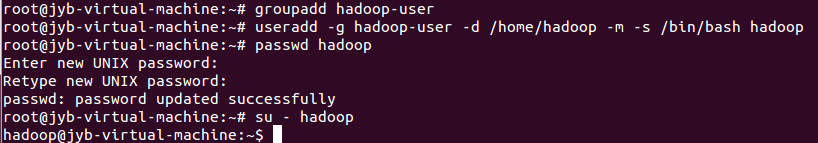
1. 即使将hadoop\_user改成hadoop-user，也会出现：

C:\Users\dell\Desktop\hadoop_exp\no_directory_with_home.png

1. 原因是没有在home目录下创建同名的文件夹，但是创建后会出现：

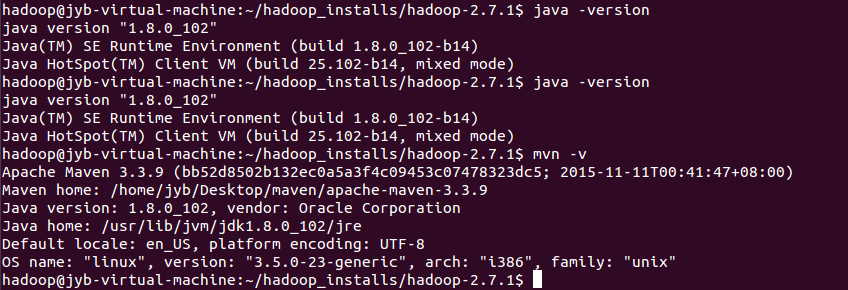


1. 这里是因为useradd命令只增加条目，而不创建用户组、指定shell版本，所以需要手动操作：



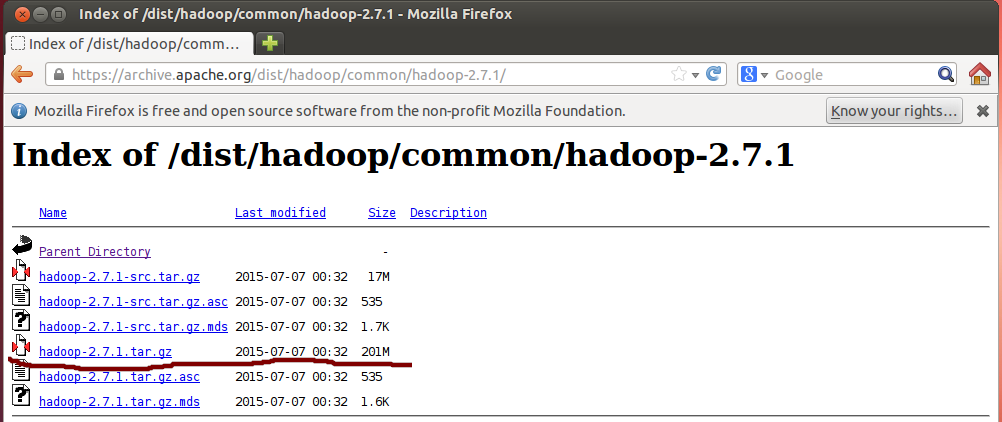
1. Java 安装与Maven安装

由于Java、Maven和IntelliJ之前已经安装，所以这里只列出版本：

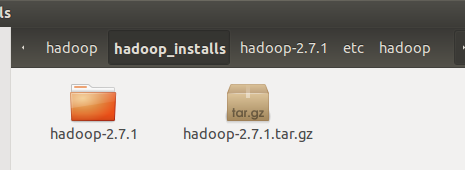


1. Hadoop 2.7.1安装

从<http://archive.apache.org/dist/hadoop/common/hadoop-2.7.1>下载源码

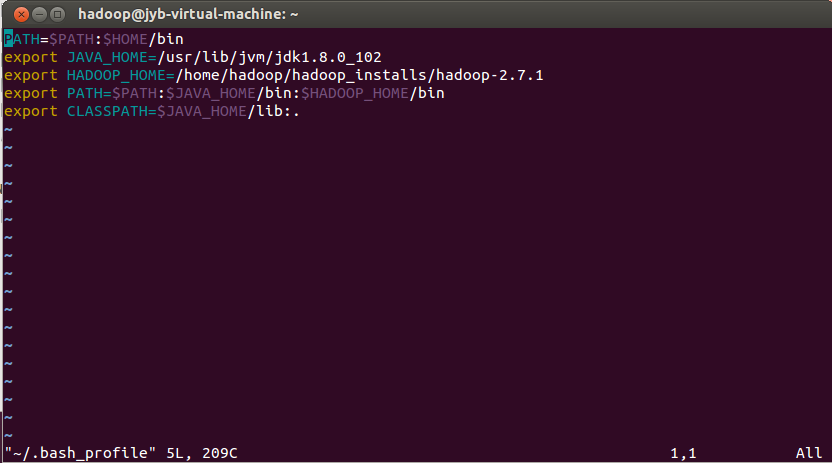


将Hadoop 2.7.1 移动到hadoop用户目录下，并解压



1. 配置Hadoop

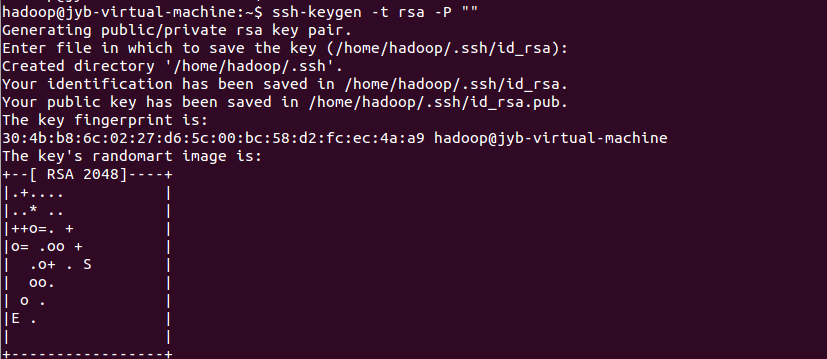
在hadoop用户的 ~/.bash\_profile配置（注，之后还会进行扩展，这里并非最终版）：



1. 安装ssh

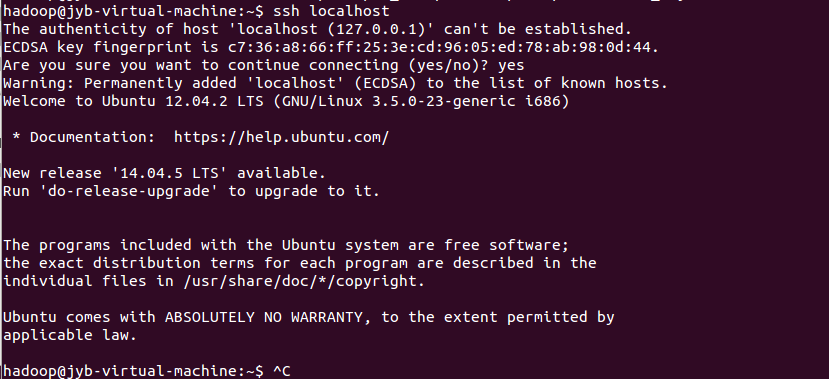
C:\Users\dell\Desktop\hadoop_exp\bad_ssh.png

这里又是PPT上的一个错误，在ssh 和-keygen之间多加了空格，应该是：

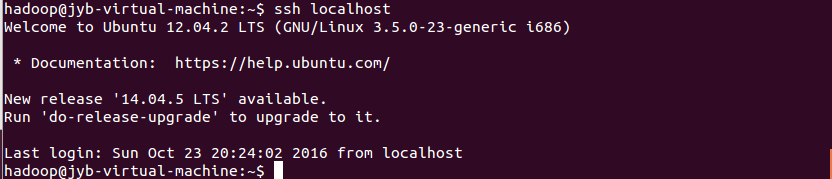


1. ssh免密码登录

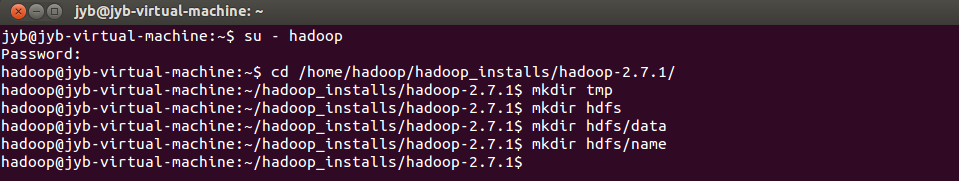
第一次使用ssh localhost会遇到如下问题



但是在第一次yes确认后，之后使用ssh localhost就不会产生问题：

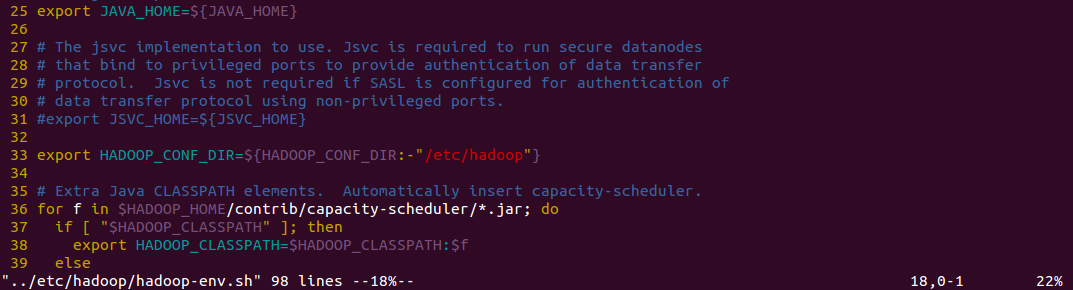


1. 在Hadoop文件夹下新建文件夹

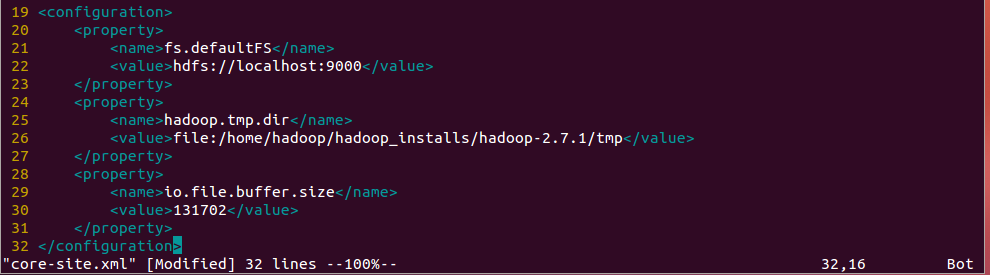


1. 配置etc/hadoop/下的文件
2. Hadoop-env.sh

hadoop-env无需配置



1. core-site.xml



1. hdfs-site.xml

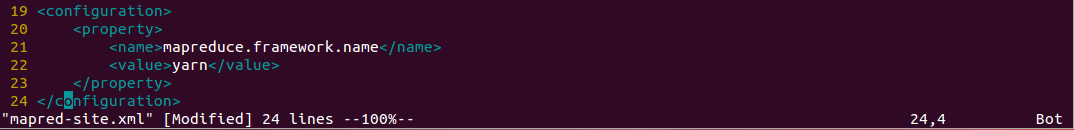


1. mapred-site.xml

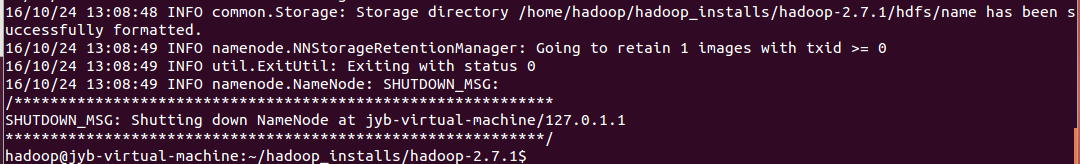
先创建mapred-site.xml

C:\Users\dell\Desktop\hadoop_exp\new_mapred-site.png

再配置



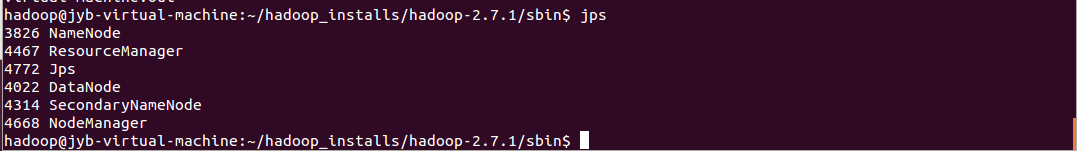
1. 由于这里没有其它主从机器，所以不需要设置master和slaves
2. 格式化Hadoop HDFS



1. 开始所有

C:\Users\dell\Desktop\hadoop_exp\start_all.png

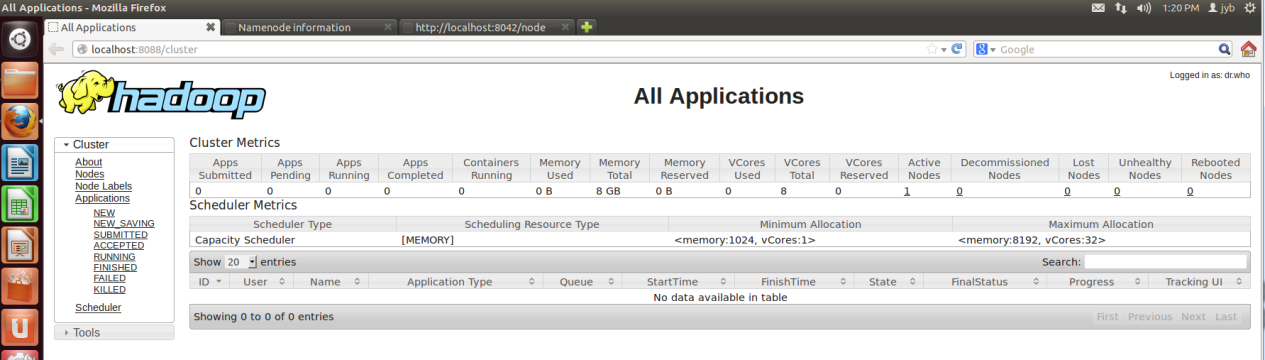
jps查看进程



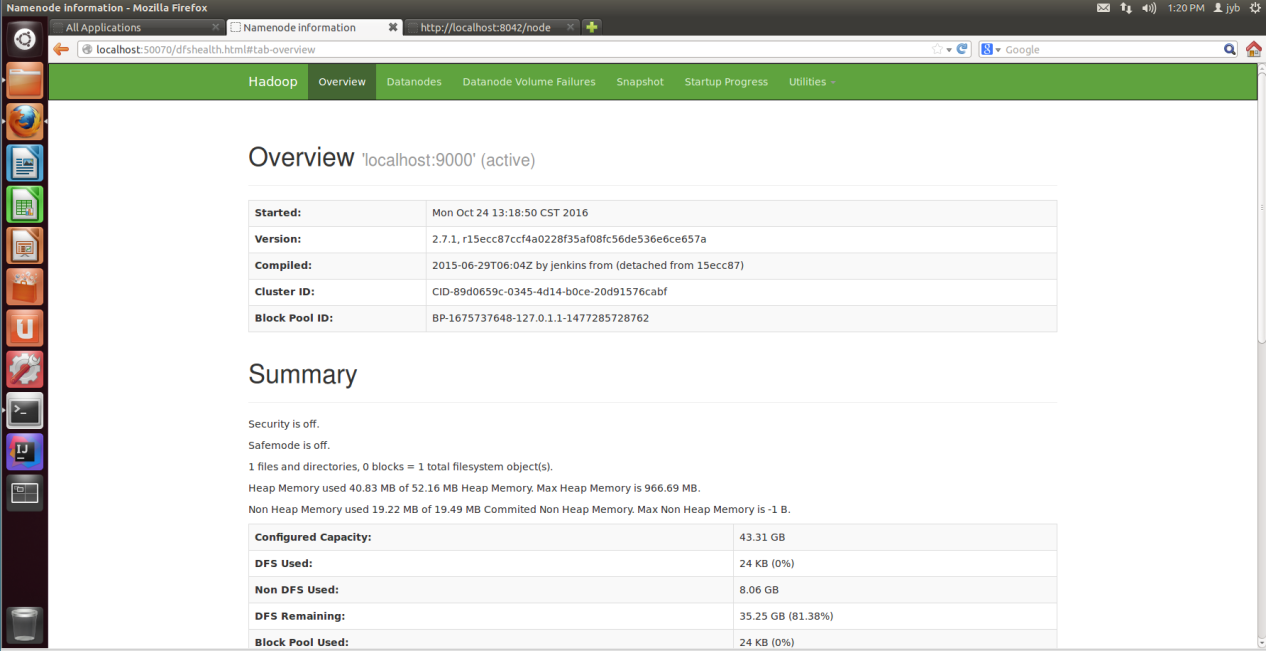
1. Hadoop Web信息

这里主机都是localhost，但是端口不一样

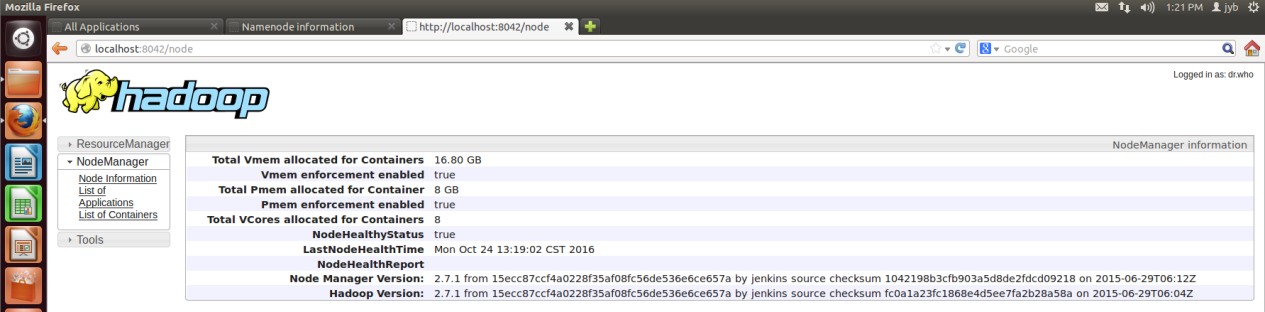
集群总览：



Namenode：

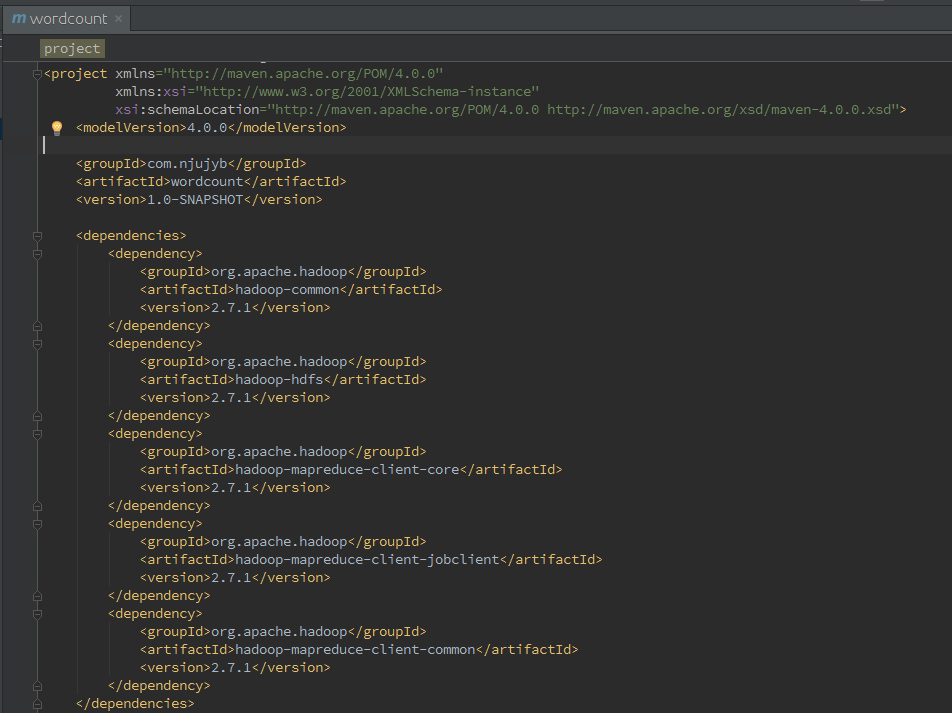


HDFS的Namenode：



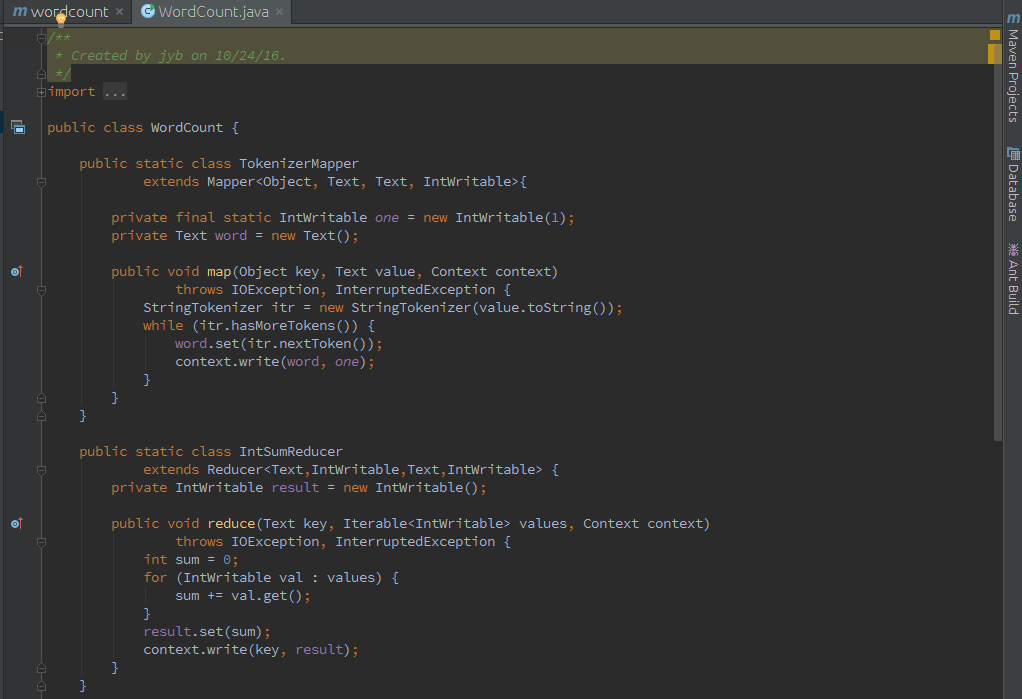
1. 用maven创建项目工程

pom.xml的配置：

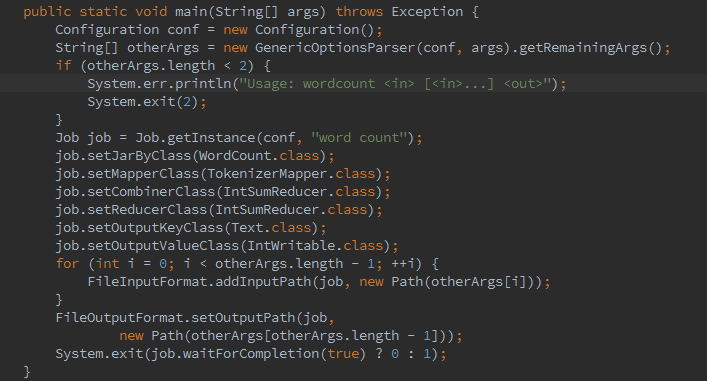


1. WordCount代码实现

Map以及Reduce代码实现：

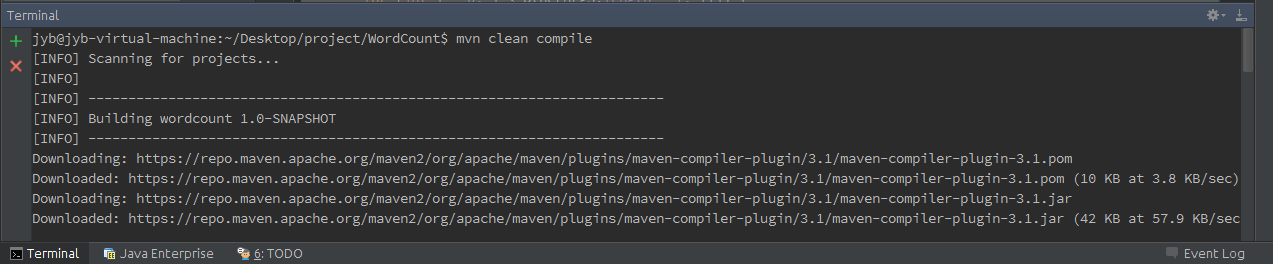


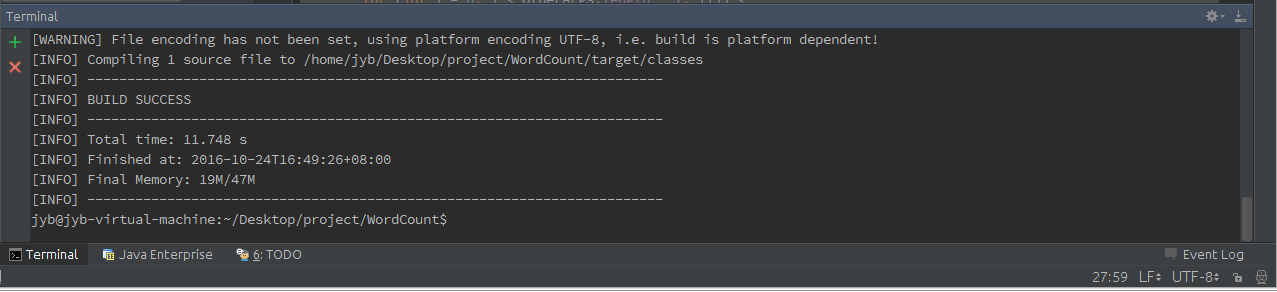
Main函数实现：



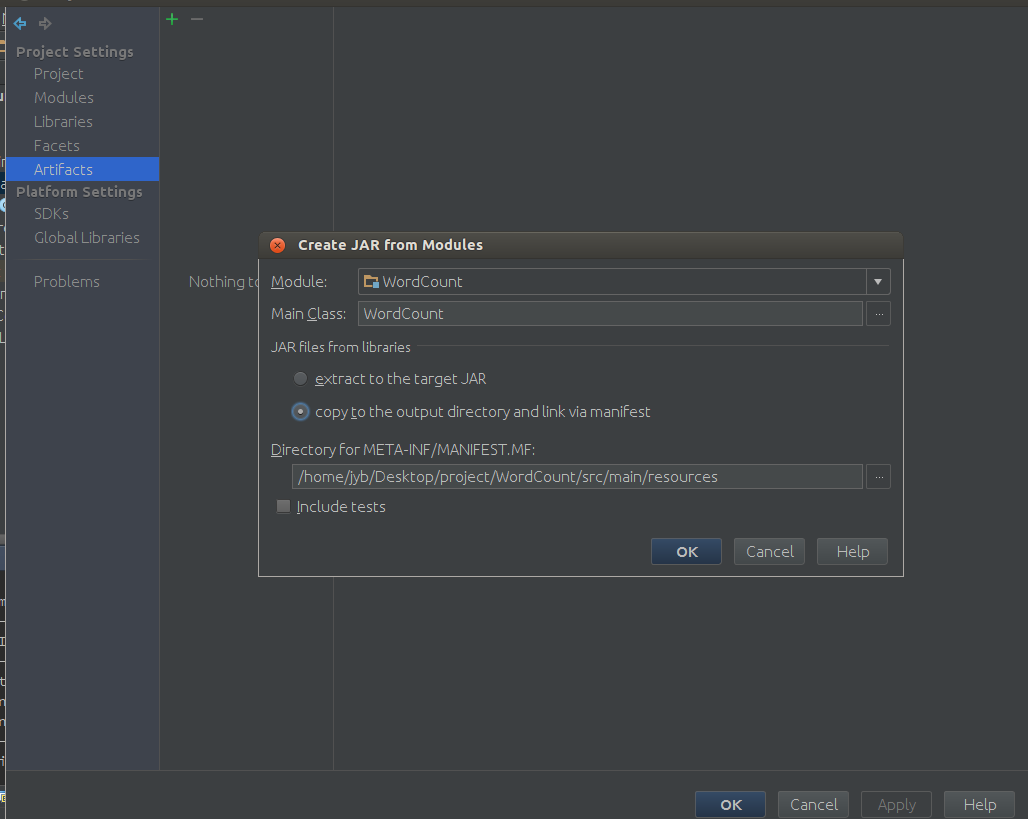
1. mvn 编译及jar 打包

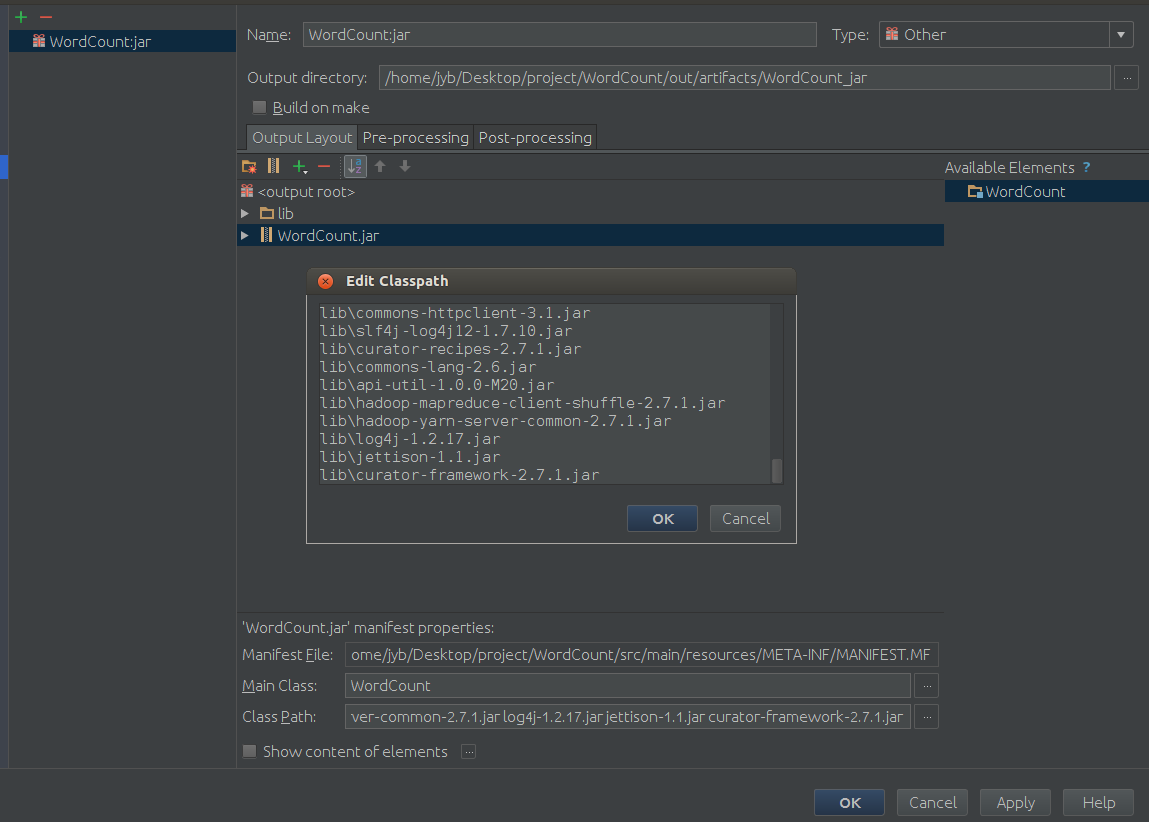
编译：



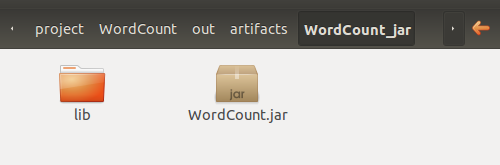


进行jar打包：



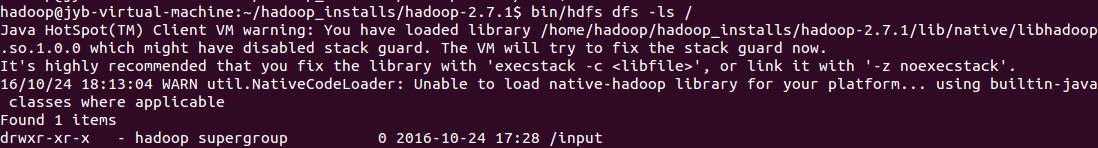


点击IntelliJ的Build进行创建：

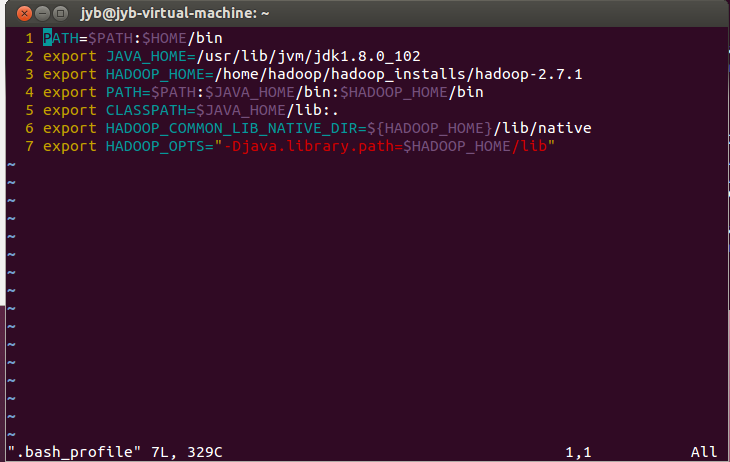


1. Hadoop编译文件与系统不一致

这个问题是因为Hadoop是在32位系统下编译的，可能与64位系统不一致



需要对~/.bash\_profile进行修改：

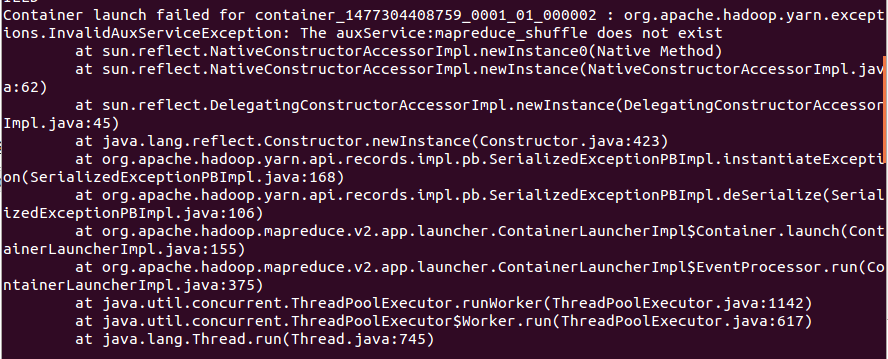


运行时仍会产生warn，不过并不影响：

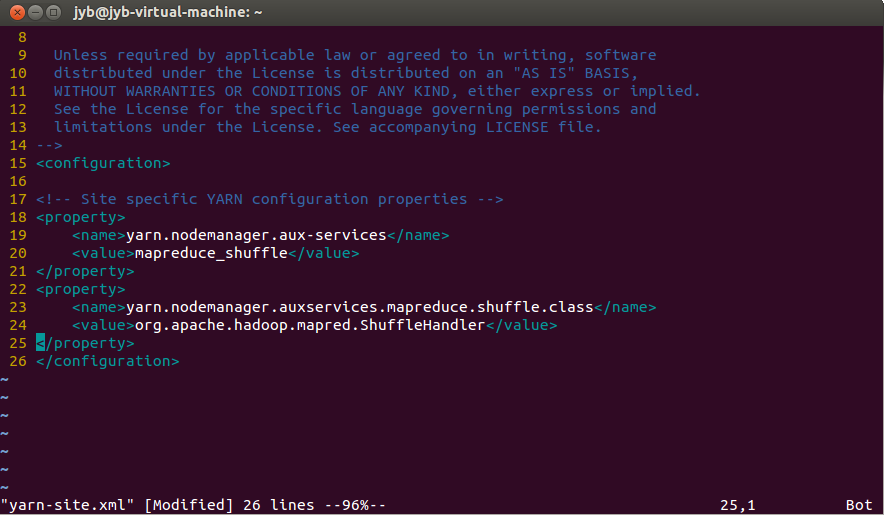


1. 运行时刻出错

这个问题是由于没有配置etc/yarn-site.xml的原因：

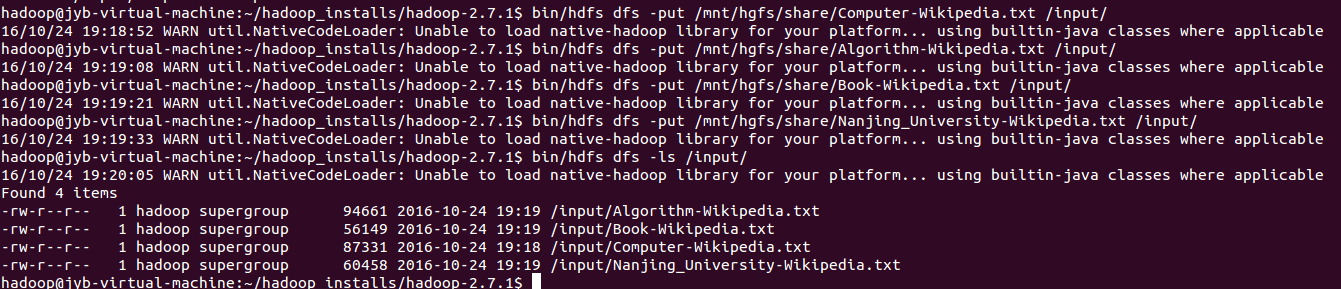


需要进行配置：



1. 输入数据

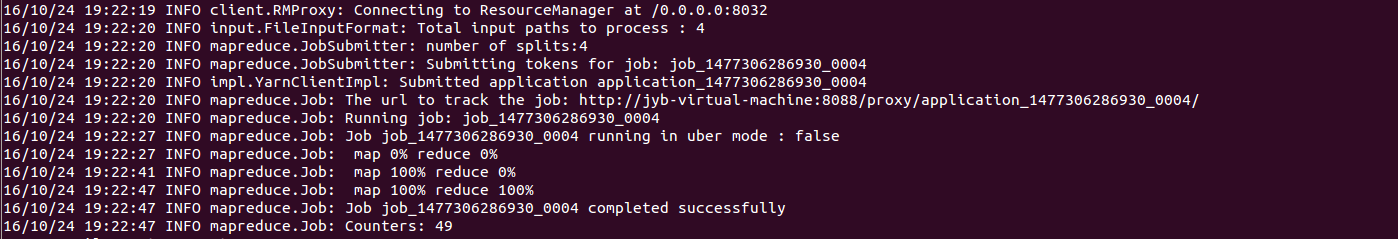
将输入数据放在HDFS下

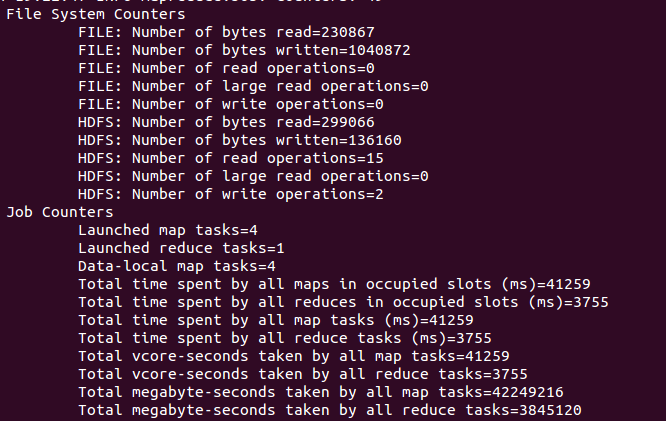


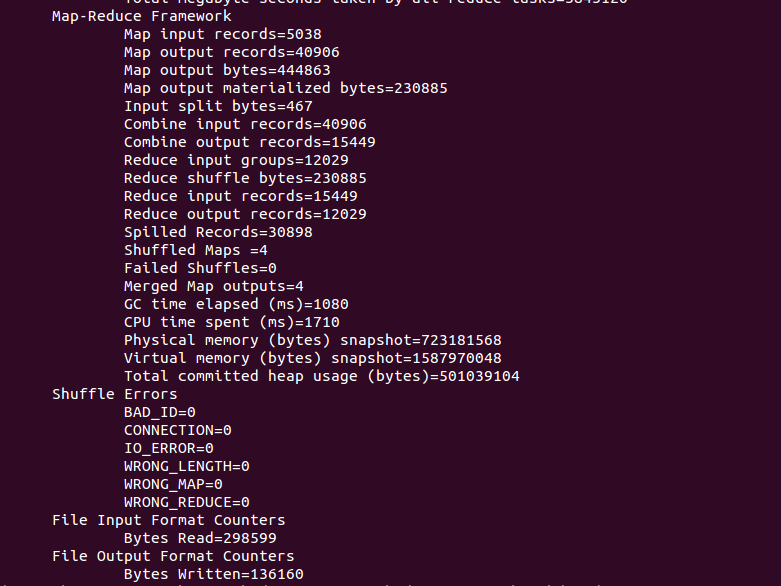
输入运行指令：

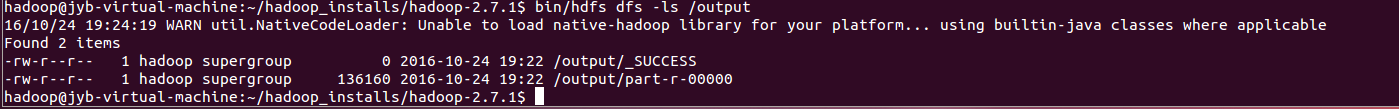
C:\Users\dell\Desktop\hadoop_exp\ready_start.png

1. 输出结果









1. 实验数据说明

数据均来自维基百科：Computer、Algorithm、Book和Nanjing University词条

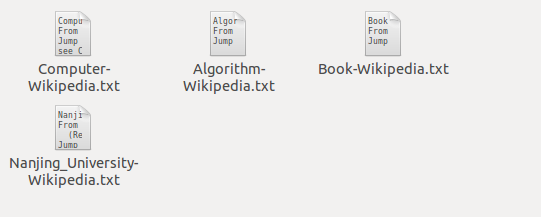
<https://en.wikipedia.org/wiki/Computer>

<https://en.wikipedia.org/wiki/Algorithm>

<https://en.wikipedia.org/wiki/Book>

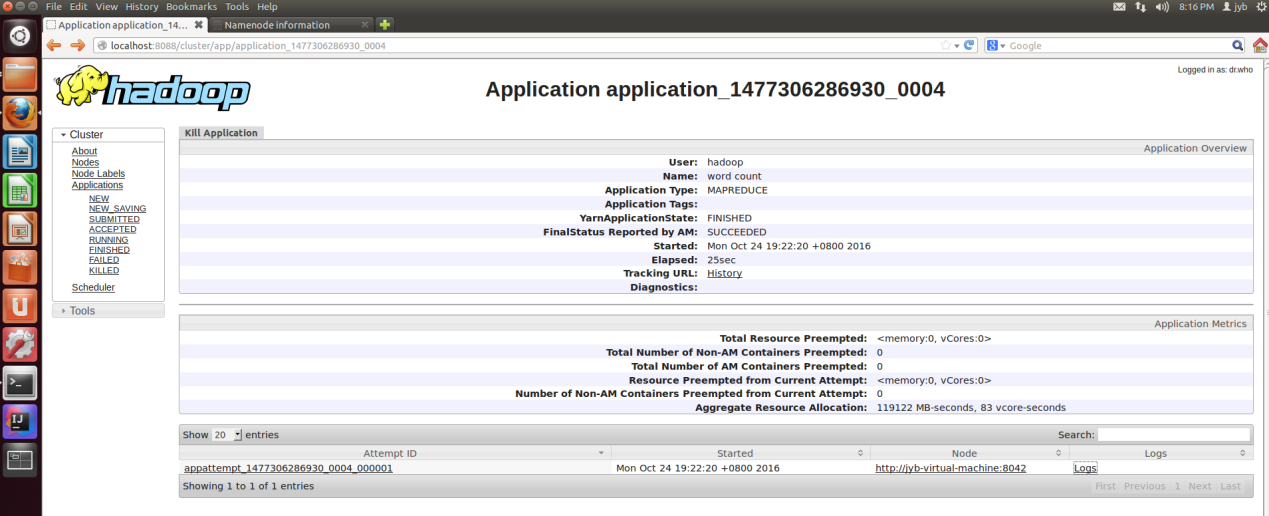
<https://en.wikipedia.org/wiki/Nanjing_university>

得到一组.txt文件：



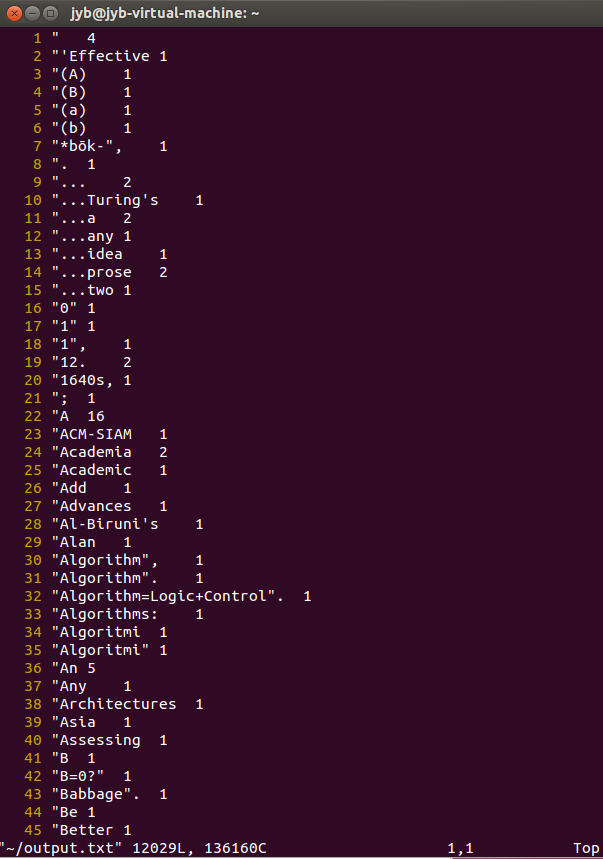
总大小298599 Bytes

1. Hadoop Web作业状态



1. 输出文件

C:\Users\dell\Desktop\hadoop_exp\output_to_file.png



1. 实验体会

Hadoop安装并不是一个困难的过程，但是需要很多细节上的注意。比如版本问题，Hadoop不同版本的文件目录格式不一样，有的是etc有的是conf等。再比如随意从网上贴的配置，往往不符合自己的实际，需要根据自己的系统反复尝试，多问问百度，就能得到正确的结果。