****

**研究生（数据中心技术课程）报告**

**题 目：Hadoop分布式实验报告**

**学 号 M20177323 8**

**姓 名 杨 呈**

**专 业 计算机技术**

**指 导 教 师**

**院（系、所） 计算机科学与技术学院**

**华中科技大学研究生院制**

**Hadoop实验报告**

一，实验名称：

hadoop分布式环境搭建和简单实例运行

二，实验目的及要求

老师您好，我们小组的论文题目是Scaling Distributed File systems In Resource-harvesting Datacenters，论文主要涉及的分布式文件系统HDFS的高扩展性研究即HD-HDFS。但是因为实验条件有限（无法搭建足够的分布式节点环境），外加之论文涉及的实验难度较大，需要对于hadoop有很深的了解才能进一步对其框架进行改变和设计。在多方思考之后，决定本着学习的态度在现有环境的基础上模拟如下实验内容：

1. hadoop分布式环境的构建。（实验结果显示当前系统时间）
2. mapreduce算法的工作原理及简单的wordcount实例运行（实验结果显示当前系统时间）。

三，实验环境：

操作系统：ubuntu16.4.0，

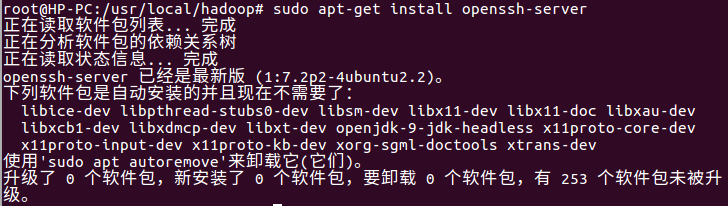
JDK版本 ：jdk-9

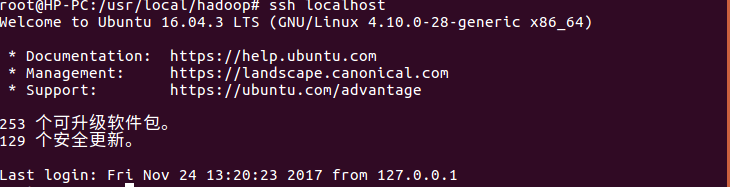
Hadoop版本：2.6.5

四，实验内容及步骤

1,安装ssh，并配置ssh自动登录

SSH全称Secure SHell，顾名思义就是非常安全的shell的意思，SSH协议是IETF（Internet Engineering Task Force）的Network Working Group所制定的一种协议。在hadoop集群中，访问其他节点需要使用ssh协议。

安装后，执行ssh localhost命令，并得到如下结果，表示安装ssh成功。



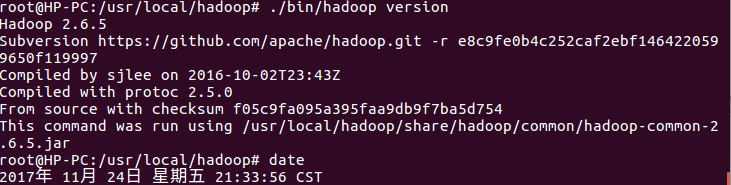
2,安装jdk及配置java环境变量

在终端执行 sudo apt-get install openjdk-9-jre-headless.之后在 执行 gedit ~/.bashrc打开文本编辑添加 JAVA\_HOME环境变量。完后后输入如下内容可获得java的安装目录。



3,安装hadoop

在hadoop官网下载压缩文件，之后在终端执行命令解压到 usr/local/hadoop 文件。执行如下命令可获得hadoop版本信息，表示hadoop安装完成

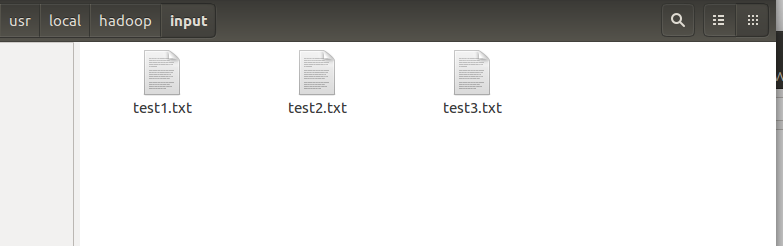


4,准备测试数据集

在hadoop目录下新建目录input用于存放测试数据集文件，用于mapreduce程序的输入

在input目录下新建测试数据三个txt文本。三个文件的内容是一致的，均为如下内容：

hello hadoop  
hello World  
Hello Java  
Hey man  
i am a programmer

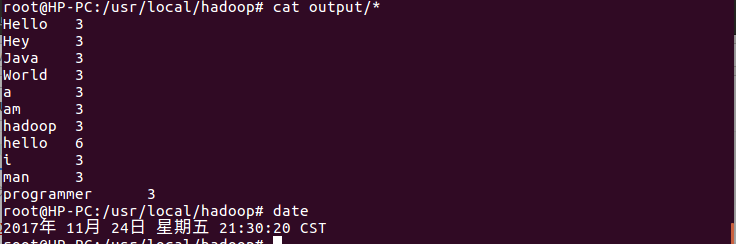


6，mapreduce程序进行单词统计

将input目录下所有文件都作为输入数据执行hadoop的mapreduce程序进行单词统计

 输出output目录下的结果统计

统计结果符合我们输入的测试数据 集内容。

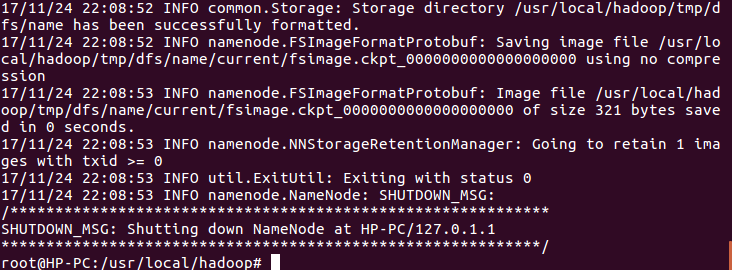
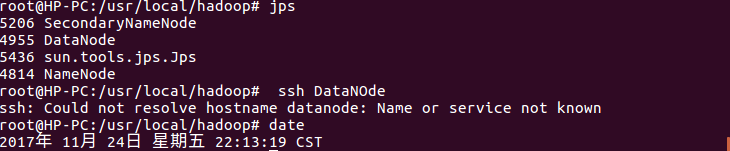


7，搭建hadoop分布式环境

初始化文件系统 HDFS

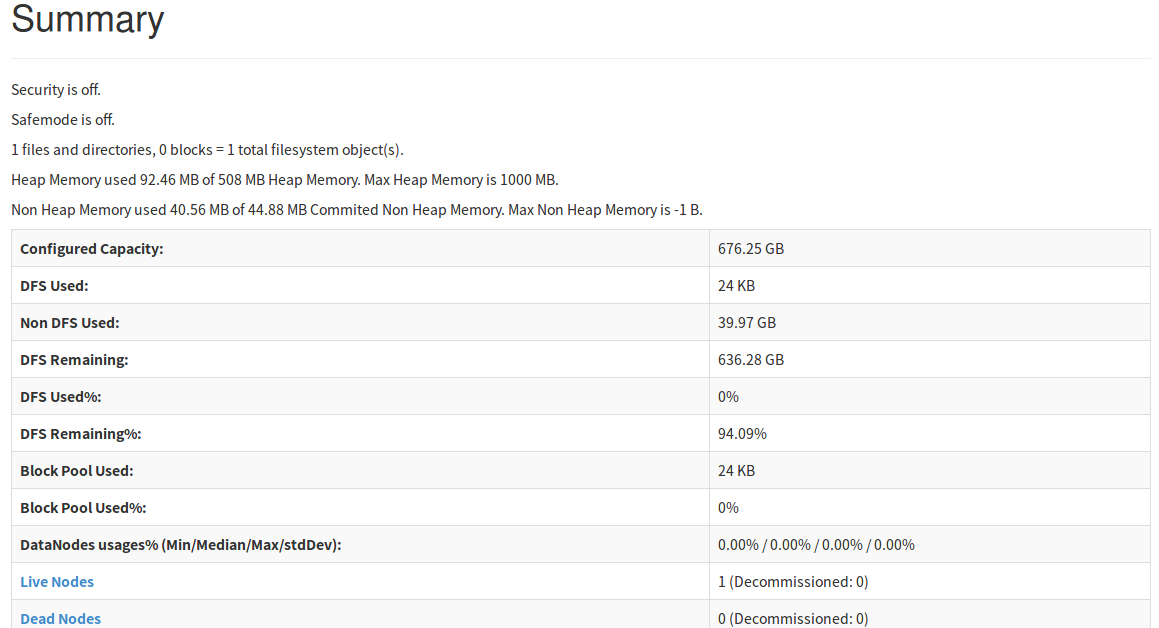


得到如下信息：

打印出hadoop分布式环境的所有节点信息。该集群包含1个namenode和一个备用namenode以及一个datanode。

8，查看分布式集群中的节点信息

在网站上显示namenode和datanode的具体信息[http://localhost:50070](http://localhost:50070/)



五，算法及原理分析

上述实例在hadoop环境下进行了简单wordcount的实验，下面将对mapreduce的原理进行简单分析在运行一个mapreduce计算任务时候，任务过程被分为两个阶段：map阶段和reduce阶段，每个阶段都是用键值对（key/value）作为输入（input）和输出（output）。而程序员要做的就是定义好这两个阶段的函数：map函数和reduce函数。

mapreduce：是一种线性的可伸缩编程模型，处理和生成超大数据集的算法模型。基本思想包括两个过程：map阶段------创建一个map函数处理一个基于key/value对的数据集，输出中间数据集合

reduce阶段-----创建一个reduce函数来合并处理中间数据内具有相同key值的value。

客户端（client）：编写mapreduce程序，配置作业，提交作业，这就是程序员完成的工作；

1. JobTracker：初始化作业，分配作业，与TaskTracker通信，协调整个作业的执行；
2. TaskTracker：保持与JobTracker的通信，在分配的数据片段上执行Map或Reduce任务，TaskTracker和JobTracker的不同有个很重要的方面，就是在执行任务时候TaskTracker可以有n多个，JobTracker则只会有一个（JobTracker只能有一个就和hdfs里namenode一样存在单点故障，我会在后面的mapreduce的相关问题里讲到这个问题的）
3. Hdfs：保存作业的数据、配置信息等等，最后的结果也是保存在hdfs上面

整个MapReduce的过程大致分为 Map-->Shuffle（排序）-->Combine（组合）-->Reduce。

map函数如下：

  protected void map(LongWritable key, Text value, Mapper<LongWritable, Text, Text, IntWritable>.Context context)

throws IOException, InterruptedException {

//默认的map的value是每一行,我这里自定义的是以空格分割

String[] vs = value.toString().split("\\s");

for (String v : vs) {

//写出去

context.write(new Text(v), ONE);

}

}

reduce函数如下：

protected void reduce(Text key, Iterable<IntWritable> values,     Reducer<Text, IntWritable, Text, IntWritable>.Context context) throws IOException, InterruptedException {

int count=0;

for(IntWritable v:values){

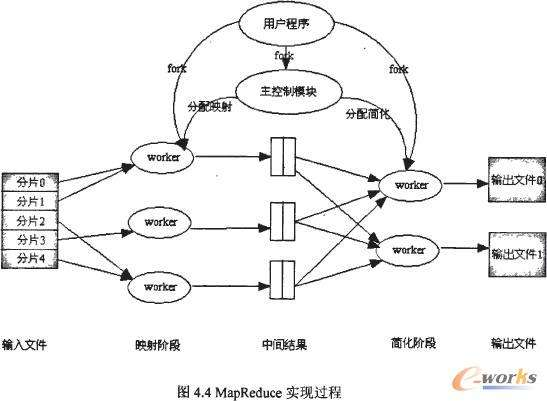
count+=v.get();//单词个数加一

}

context.write(key, new IntWritable(count));

}

Mapreduce实现过程图如下所式：



六，实验结果及评估

通过实验和原理的学习以及操作，深入了解hadoop分布式结构以及mapreduce的工作原理，为进一步学习大数据的其他框架技术打下基础。同时在linux系统上运行，加强了对于linux操作系统的熟悉，并了解常用shell命令。

实验初步完成了Hadoop分布式环境的构建并运行了简单实例，达到了预期目标，虽然无法进一步完成对于论文研究的实验操作，但是通过基础的实验足以初步了解数据中心扩展方面的知识和理论，加深对于整个hdfs分布式文件系统理论和原理的理解。