

山东大学 计算机科学与技术 学院  
云计算技术 课程实验报告

学号：201900130133	姓名：施政良	班级：四班
实验题目：面向分布式存储和计算的 Hadoop 练习		
实验学时：2	实验日期：2020-04-18	
<p><b>实验目的：</b>在 Linux 环境下，熟悉 Hadoop 环境。</p> <p><b>具体包括：</b>了解 Hadoop 生态结构和关键工具/组件，了解 Hadoop 编程环境的配置和部署，完成实验环境及实验工具的熟悉，撰写实验报告。</p>		
<p><b>硬件环境：</b></p> <p>联网计算机一台</p>		
<p><b>软件环境：</b></p> <p>Windows or Linux</p>		
<p><b>实验步骤与内容：</b></p> <p><b>实验步骤概述：</b></p> <p>本次试验旨在介绍 Hadoop 的基本概念，需要了解 Hadoop 生态结构和关键工具/组件。同时需要熟悉 Linux 下 Hadoop 的开发环境，完成实验环境及实验工具的配置。</p> <p>具体实验步骤可以划分为如下几个步骤</p> <ol style="list-style-type: none"><li>1. Hadoop 的介绍</li><li>2. Hadoop 的组成元素</li><li>3. Hadoop 的环境配置</li><li>4. 实验总结与体会</li></ol> <p><b>具体实验内容</b></p> <p>一、Hadoop 介绍</p> <p>Apache Hadoop 是一款支持数据密集型分布式应用程序并以 Apache 2.0 许可协议发布的开源软件框架。它支持在商用硬件构建的大型集群上运行的应用程序。Hadoop 是根据谷歌公司发表的 MapReduce 和 Google 档案系统的论文自行实作而</p>		

成。所有的 Hadoop 模块都有一个基本假设，即硬件故障是常见情况，应该由框架自动处理。

Hadoop 框架透明地为应用提供可靠性和数据移动。它实现了名为 MapReduce 的编程范式：应用程序被分割成许多小部分，而每个部分都能在集群中的任意节点上执行或重新执行。此外，Hadoop 还提供了分布式文件系统，用以存储所有计算节点的数据，这为整个集群带来了非常高的带宽。MapReduce 和分布式文件系统的设计，使得整个框架能够自动处理节点故障。它使应用程序与成千上万的独立计算的电脑和 PB 级的数据连接起来。现在普遍认为整个 Apache Hadoop “平台”包括 Hadoop 内核、MapReduce、Hadoop 分布式文件系统（HDFS）以及一些相关项目，有 Apache Hive 和 Apache HBase 等等

## 2. Hadoop 的组成架构

Hadoop 由许多元素构成。其最底部是 Hadoop Distributed File System（HDFS），它存储 Hadoop 集群中所有存储节点上的文件。HDFS 的上一层是 MapReduce 引擎，该引擎由 JobTrackers 和 TaskTrackers 组成。通过对 Hadoop 分布式计算平台最核心的分布式文件系统 HDFS、MapReduce 处理过程，以及数据仓库工具 Hive 和分布式数据库 Hbase 的介绍，基本涵盖了 Hadoop 分布式平台的所有技术核心

## 3. Hadoop 环境配置

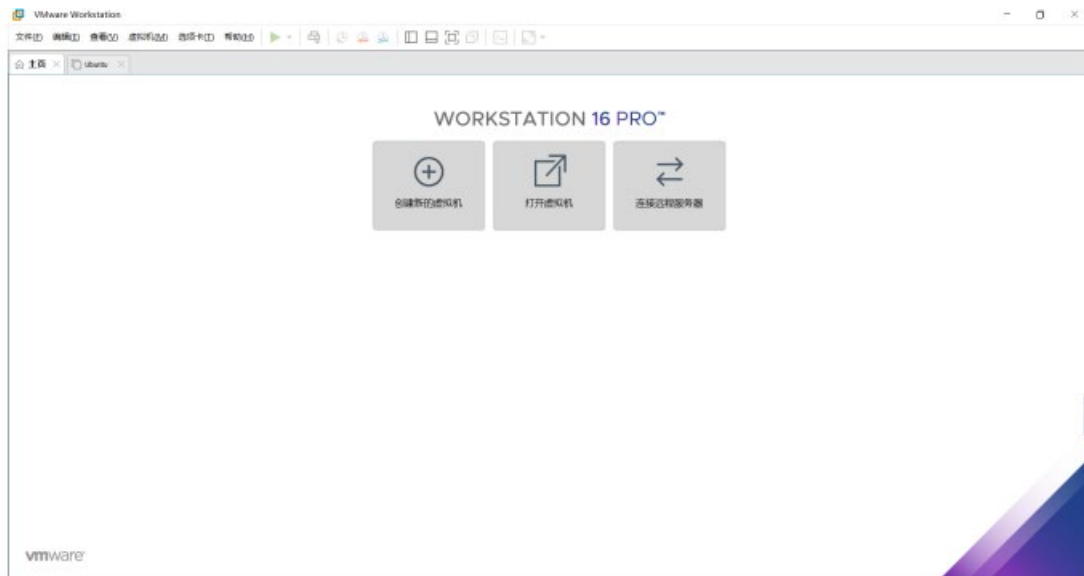
### 3.1 安装对应版本的虚拟机

由于在之前的实验中使用 Ubuntu18.04 作为实验环境，因此在本次试验中首先安装 Ubuntu20.04 作为 Linux 实验环境。

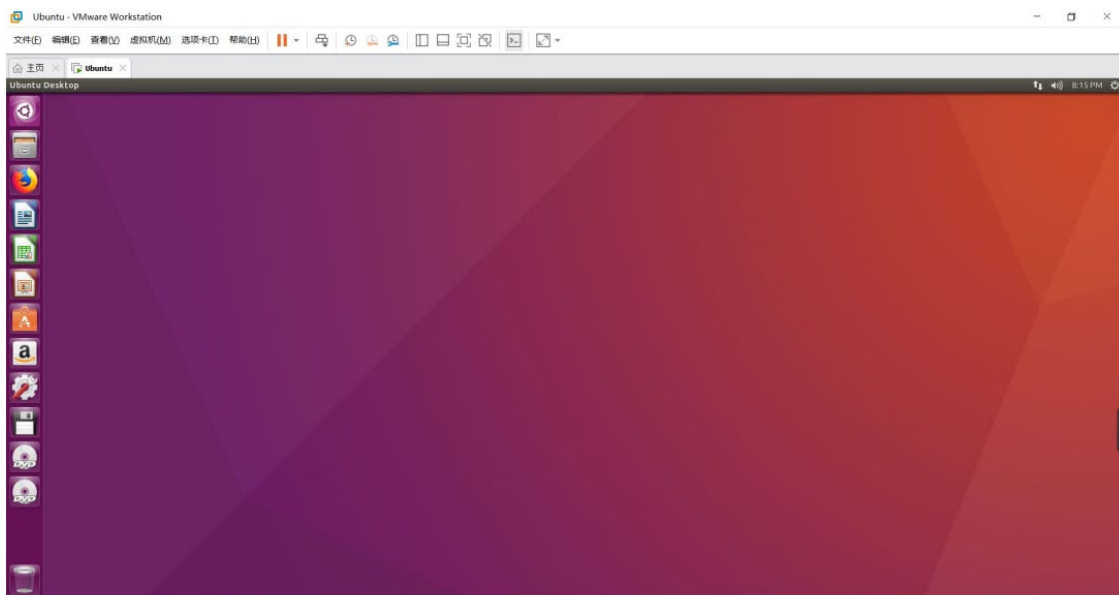
同实验一，在本次实验中采用 VMware 虚拟机，并采用 ubuntu 配置 linux 环境。Ubuntu 一个以桌面应用为主的 Linux 操作系统其界面与常用的 Windows 界面非常相似。解决了 Linux 安装和使用困难的问题，同时，其通过

`sudo` 指令执行系统相关的任务的设置也 使其比传统的以系统管理员账号进行管理工作的方式更为安全。

打开 `vmware` 并创建虚拟机



将 `ubuntu` 镜像文件装入虚拟机中并选择合适的硬件配置（如磁盘大小，内核数量） 即可。最终完成虚拟机的创建，如下图所示。



### 3.2 创建 `hadoop` 用户

在终端中输入如下命令创建 Hadoop 用户

1. 为 `hadoop` 用户设置密码: `sudo passwd hadoop`

2. 为 `hadoop` 用户增加管理员权限：`sudo adduser hadoop sudo` #
3. 注销当前用户，并使用 `hadoop` 用户登录：`su - hadoop`

### 3.3 安装 JDK

JDK 为 `java` 运行的必要环境。在本次实验中，首选创建 `java` 目录，例如#创建 `jvm` 文件夹

```
mkdir /usr/lib/jvm
```

解压到目录下

```
sudo tar zxvf jdk-18_linux-aarch64_bin.tar.gz -C /usr/lib/jvm
```

其中 `sudo` 为 `root` 权限, `tar` 为解压命令, `zxvf` 为 `tar` 的命令行参数, `jdk-18_linux-aarch64_bin.tar.gz` 是 `jdk` 压缩包的文件名

上述指令将 `JDK` 解压到 `/usr/lib/jvm` 目录下。

之后进入该目录

```
cd /usr/lib/jvm
```

为了便于后续实验的进行，此处使用 `mv` 指令，将文件夹重命名为 `java` 文件名规范。

### 3.4 配置 `java` 环境变量

在终端使用 `vim` 编辑器对根目录下的 `./bashrc` 文件进行编辑，添加相应的文件路径

```
vim ~/.bashrc
```

在 `~/.bashrc` 最后添加下列代码并保存

```
#Java Environment
export JAVA_HOME=/usr/lib/jvm/java
export JRE_HOME=${JAVA_HOME}/jre
export CLASSPATH=.:${JAVA_HOME}/lib:${JRE_HOME}/lib
export PATH=${JAVA_HOME}/bin:$PATH
```

### 3.5 更新配置，并测试是否安装成功

配置 java 环境之后，需要更新并测试安装是否成功。使用如下命令使新配置的环境变量生效

```
source ~/.bashrc
```

打印 Java 版本，测是否安装成功

```
java -version
```

如下图所示

```
java version "1.8.0_321"  
Java(TM) SE Runtime Environment (build 1.8.0_321-b07)  
Java HotSpot(TM) 64-Bit Server VM (build 25.321-b07, mixed mode)
```

### 3.5. 安装 ssh

首先安装 SSH server

```
sudo apt-get install openssh-server
```

之后登录本机测试，需要手动输入 “yes”

```
ssh localhost
```

过程如下所示

```
ECDSA key fingerprint is SHA256:UdcjHkY3V/V4sV4Ypj2RRlwEUEC8rxzErJ3IDRvuoIO.  
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes  
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.  
hadoop@localhost's password:  
Welcome to Ubuntu 20.04.4 LTS (GNU/Linux 5.13.0-30-generic x86_64)  
  
* Documentation:  https://help.ubuntu.com  
* Management:    https://landscape.canonical.com  
* Support:        https://ubuntu.com/advantage  
  
0 updates can be applied immediately.  
  
Your Hardware Enablement Stack (HWE) is supported until April 2025.  
  
The programs included with the Ubuntu system are free software;  
the exact distribution terms for each program are described in the  
individual files in /usr/share/doc/*/copyright.
```

### 3.6 安装单机 Hadoop

解压安装 Hadoop 到/usr/local 目录下

之后进入目录

```
cd /usr/local
```

配置环境变量，使用 vim 在 ~/.bashrc 中添加如下代码并保存

```
#Hadoop Environment
export HADOOP_HOME=/usr/local/hadoop
export CLASSPATH=$(HADOOP_HOME/bin/hadoop classpath):$CLASSPATH
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

刷新 ~/.bashrc 配置文件

```
source ~/.bashrc
```

之后输入 `hadoop version`，测试是否安装成功

```
hadoop@ubuntu: /usr/local$ hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3.2.2.jar
```

### 3.7 伪分布式 Hadoop

在 /usr/local/hadoop/etc/hadoop 目录 `hadoop-env.sh` 添加 Java 路径

```
export JAVA_HOME=/usr/lib/jvm/java
```

修改配置文件 `core-site.xml` Hadoop 的配置文件位于 /usr/local/hadoop/etc/hadoop/ 中修改 `core-site.xml` 文件，添加下列内容。

```
<configuration>
  <property>
    <name>hadoop.tmp.dir
    </name>
    <value> file:/usr/local/hadoop/tmp</value>
    <description>Abase for other temporary directories.</description>
```

```
    </property>
    <property>
      <name>fs.defaultFS</name>
      <value>hdfs://localhost:9000</value>
    </property>
  </configuration>
```

修改配置文件 hdfs-site.xml 添加下列内容，

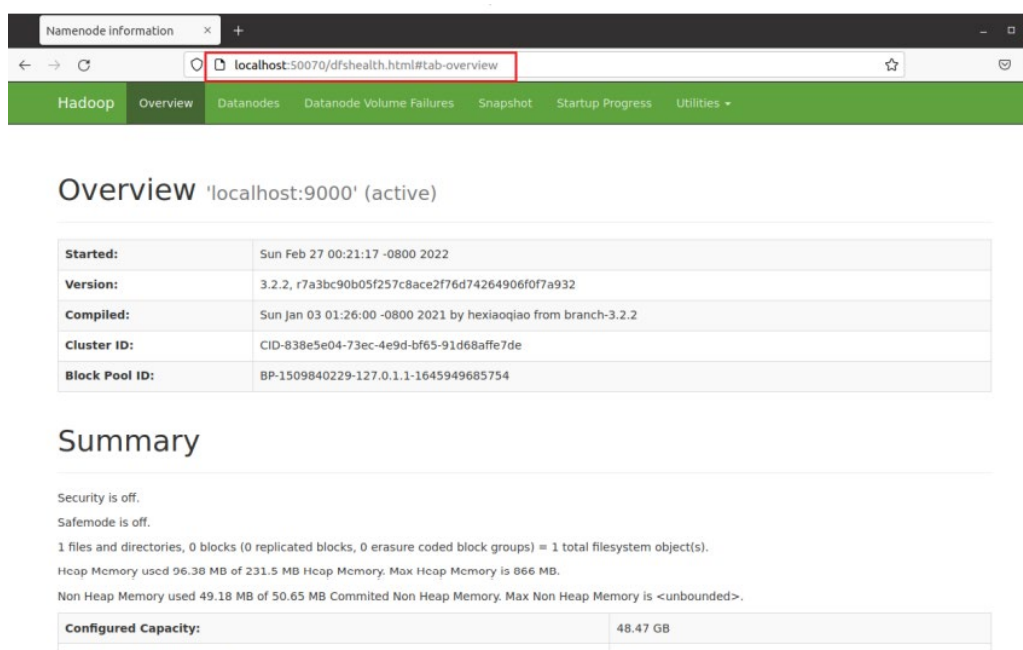
```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/usr/local/hadoop/tmp/dfs/name</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/usr/local/hadoop/tmp/dfs/data</value>
  </property>
  <property>
    <name>dfs.http.address</name>
    <value>0.0.0.0:50070</value>
  </property>
</configuration>
```

之后格式化集群节点 `hdfs namenode -format`，并启动 `hadoop`

`start-dfs.sh`

```
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu]
ubuntu: Warning: Permanently added 'ubuntu' (ECDSA) to the list of known hosts.
2022-02-27 00:17:10,705 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

使用 `jps` 查看进程，出现 `DataNode`, `NameNode`, `SecondaryNameNode` 即启动成功。在浏览器输入 `localhost:50070` 查看 `hadoop` 状态。



The screenshot shows a web browser window with the address bar displaying `localhost:50070/dfshealth.html#tab-overview`. The page title is "Namenode information". The navigation bar includes "Hadoop", "Overview", "Datanodes", "Datanode Volume Failures", "Snapshot", "Startup Progress", and "Utilities". The main content area is titled "Overview 'localhost:9000' (active)". It contains a table with the following information:

Started:	Sun Feb 27 00:21:17 -0800 2022
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled:	Sun Jan 03 01:26:00 -0800 2021 by hexiaoqiao from branch-3.2.2
Cluster ID:	CID-838e5e04-73ec-4e9d-bf65-91d68affe7de
Block Pool ID:	BP-1509840229-127.0.1.1-1645949685754

Below the table is a "Summary" section with the following text:

Security is off.  
Safemode is off.  
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).  
Heap Memory used 96.38 MB of 231.5 MB Heap Memory. Max Heap Memory is 866 MB.  
Non Heap Memory used 49.18 MB of 50.65 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

At the bottom, there is a table showing "Configured Capacity: 48.47 GB".

关闭 Hadoop `stop-dfs.sh`

结论分析与体会：

## 1. Hadoop 的优点

Hadoop 是一个能够对大量数据进行分布式处理的软件框架。Hadoop 以一种可靠、高效、可伸缩的方式进行数据处理。

Hadoop 是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。

Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。

Hadoop 还是可伸缩的，能够处理 PB 级数据。此外，Hadoop 依赖于社区服务，因此它的成本比较低，任何人都可以使用。



Hadoop 是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在 Hadoop 上开发和运行处理海量数据的应用程序。

它主要有以下几个优点

- 高可靠性。Hadoop 按位存储和处理数据的能力值得人们信赖。
- 高扩展性。Hadoop 是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。
- 高效性。Hadoop 能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快
- 高容错性。Hadoop 能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。
- 低成本。与一体机、商用数据仓库以及 QlikView、Yonghong Z-Suite 等数据集市相比，hadoop 是开源的，项目的软件成本因此会大大降低。
- Hadoop 带有用 Java 语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。Hadoop 上的应用程序也可以使用其他语言编写，比如 C++

## 2. Hadoop 有哪些应用

Hadoop 的最常见用法之一是 Web 搜索。虽然它不是唯一的软件框架应用程序，但作为一个并行数据处理引擎，它的表现非常突出。Hadoop 最有趣的方面之一是 Map and Reduce 流程，它受到 Google 开发的启发。这个流程称为创建索引，它将 Web 爬行器检索到的文本 Web 页面作为输入，并且将这些页面上的单词的频率报告作为结果。然后可以在整个 Web 搜索过程中使用这个结果从已定义的搜索参数中识别内容

### 体会

在本次实验中，我了解了 Hadoop 的相关知识。Hadoop 得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载(ETL)方面上的天然优势。Hadoop 的分布式架构，将大数据处理引擎尽可能的靠近存储，对例如像 ETL 这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接走向存储。Hadoop 的 MapReduce 功能实现了将单个任务打碎，并将碎片任务(Map)发送到多个节点上，之后再以单个数据集的形式加载(Reduce)到数据仓库里

作为实验的总结，本次实验中通过实际配置环境，使我初步熟悉了 Hadoop 的开发环境，同时通过相关资料了解相比于其他平台，Hadoop 有高可靠性、高可扩展性、高容错性和高效性。目前 Hadoop 技术在互联网领域已经得到了广泛的运用。