大数据技术基础实验一
实验报告

毛子恒
2019211397
北京邮电大学 计算机学院

日期：2022 年 3 月 9 日

Part I

实验一（一）

1 概述

1.1 实验目的

1. 了解华为云基础操作；
2. 购买华为云 ECS；
3. 掌握华为云服务器操作。

1.2 实验步骤

1. 购买 4 个 ECS 实例；
2. 创建 OBS 桶，获取访问密钥和 endpoint。

2 实验结果及分析

购买 ECS 登录华为云控制台，依照实验指导书的步骤购买 ECS，购买结果如图 1。

名称/ID	监控	可用区	状态	规格/镜像	IP地址	计费模式	标签	操作
mzh_2019211397-0004 883d5f20-29c6-40f6-afa7-c8892a509f37		可用区2	运行中	2vCPUs | 4 GiB | kc1.large.2 CentOS 7.6 64bit with ARM	120.46.147.198 (弹性公网... 192.168.0.236 (私有)	按需计费 2022/03/08 14:25:58 创...	--	远程登录 | 更多
mzh_2019211397-0001 b5b3988a-c227-4bf6-bbf0-e8d79469f80f		可用区2	运行中	2vCPUs | 4 GiB | kc1.large.2 CentOS 7.6 64bit with ARM	120.46.147.42 (弹性公网)... 192.168.0.168 (私有)	按需计费 2022/03/08 14:25:57 创...	--	远程登录 | 更多
mzh_2019211397-0003 b8d8d958-46c1-4da5-b28a-6a16119454f8		可用区2	运行中	2vCPUs | 4 GiB | kc1.large.2 CentOS 7.6 64bit with ARM	119.3.231.14 (弹性公网) ... 192.168.0.65 (私有)	按需计费 2022/03/08 14:25:52 创...	--	远程登录 | 更多
mzh_2019211397-0002 259ce4a9-0ef6-42c3-9e25-ac92d39db1a4		可用区2	运行中	2vCPUs | 4 GiB | kc1.large.2 CentOS 7.6 64bit with ARM	124.70.67.157 (弹性公网)... 192.168.0.249 (私有)	按需计费 2022/03/08 14:25:52 创...	--	远程登录 | 更多

图 1: 购买 ECS

创建 OBS 桶 依照实验指导书的步骤创建 OBS 桶，结果如图 2。

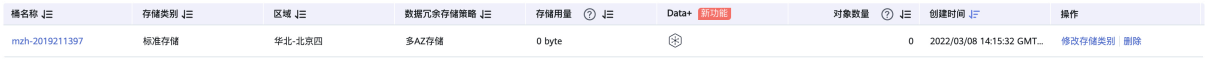

图 2: 创建 OBS 桶

获取访问密钥和 endpoint 依照实验指导书的步骤获取访问密钥和 endpoint，结果如图 3和图 4。

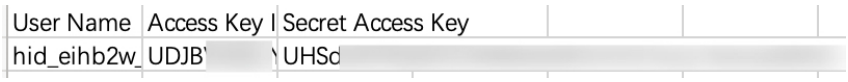

图 3: 访问密钥

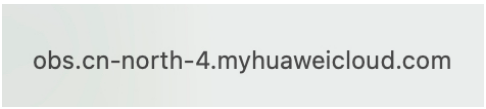

图 4: endpoint

3 实验总结

本次实验中我基本熟悉了华为云 ECS 的基本操作，创建了 ECS 和 OBS 实例，为之后的实验做了准备工作。

Part II

实验一（二）

4 概述

4.1 实验目的

1. 学习搭建 Hadoop 集群；
2. 学习创建 Maven 工程；
3. 掌握 HDFS 文件读写操作。

4.2 实验步骤

1. Hadoop 集群搭建；
2. 创建 Maven 工程；
3. Java 实现 HDFS 文件读写。

5 实验结果及分析

Hadoop 集群搭建 依照实验指导书的步骤搭建 Hadoop 集群，有以下几点需要注意：

1. 将主机名和 IP 替换成自己的主机名和 IP；
2. 执行命令或者配置环境变量时，注意删去多余的空格，比如实验指导书中，设置`hadoop-env.sh`的环境变量时，`JAVA_HOME` 的路径中有空格，需要删除。
3. 部分图片与文字描述不符，以文字为准；
4. 设置 `core-site.xml` 时，`hadoop.tmp.dir` 属性的值应该为`/home/modules/hadoop-2.7.7/tmp`。
5. `scp` 命令的前一个参数为`/home/modules/hadoop-2.7.7`，注意不要多一个`/`，否则会将子文件夹拷贝过去。
6. 配置出错和关闭服务器之前执行 `stop-all.sh`。
7. 如果 node1 中也有 DataNode，首先检查 `slaves` 文件里应该没有 node1，并且执行 `hdfs namenode -format` 且重启。

成功搭建并启动 Hadoop 集群的结果如图 5，主机有四个进程，分别是主次 NameNode、资源管理和 Jps，从机有三个进程，分别是节点管理、Jps 和 DataNode。

```
[root@mzh-2019211397-0001 ~]# jps
14141 SecondaryNameNode
13928 NameNode
14324 ResourceManager
15640 Jps
```

(a) node1

```
[root@mzh-2019211397-0002 ~]# jps
4797 NodeManager
5013 Jps
4679 DataNode
```

(b) node2

图 5: 在 node1 和 node2 中执行 jps 命令的结果

创建 Maven 工程 依照实验指导书的步骤创建 Maven 工程，注意如果是全新的环境，在创建项目时需要自行下载一个 1.8 版本的 SDK。

Java 实现 HDFS 文件读写 依照实验指导书的步骤编写 Java 代码，有以下几点需要注意：

1. 使用 IDEA 自动添加 import 时，可能会导入 Java 自定义的虚类，需要改成 Hadoop 实现的子类，如图 6。
2. 所有方法改为静态方法。
3. 对于 `log4j: WARN No appenders could be found for logger` 的报错，需要增加一个 log4j 的配置文件，参考这里。
4. 服务器需要设置安全组，开放 8020 端口进行 NameNode 的 RPC 调用，50010 端口用于 DataNode 的数据传输，还有一些其他的 Web 端口用于查看状态，具体如图 7。

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FSDataOutputStream;
import org.apache.hadoop.fs.FileStatus;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IOUtils;
```

图 6: import

图 7: 安全组

代码运行结果如图 8，首先查看 HDFS 目录为空，之后上传了一个文件，又写入了一个文件 `mzh_2019211397.txt`，将刚才写入的文件下载下来，再查看 HDFS 的根目录，现在有两个文件。

```
/Users/xqmmcqs/Library/Java/JavaVirtualMachines/corretto-1.8.0_322/Contents/Home/bin/java ...
View file:
 WARN [main] - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HDFS is empty.
Upload file:
Upload successfully!
Write file:
hdfs://120.46.147.42/mzh_2019211397.txt
Download file:
Download successfully!
View file:
name: hdfs://120.46.147.42/mzh_2019211397.txt, folder: false, size: 57
name: hdfs://120.46.147.42/upload_2019211397.txt, folder: false, size: 0
```

图 8: 运行结果

下载完成的文件如图 9，内容为我们代码中刚刚写入的。

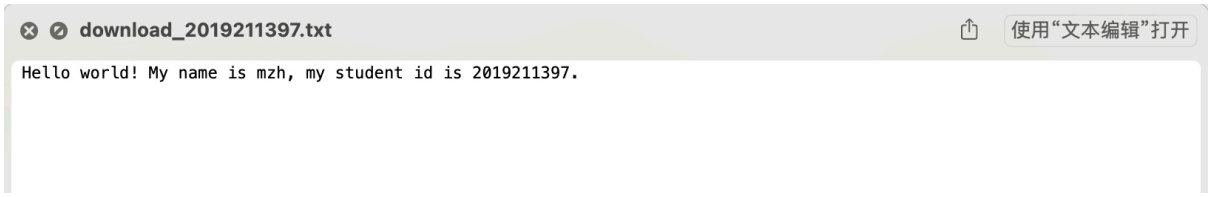

图 9: 下载的文件

6 实验总结

本次实验中我搭建了 Hadoop 集群，并且编写 Java 代码访问 HDFS，进行了基本的文件操作。经过这次实验，我对 Hadoop 基本概念和特性有了更深的体会，同时增强了信息获取能力和 Java 代码能力。