课程实验一：云主机实现大数据

实验时间：2021 年 03 月 25 日

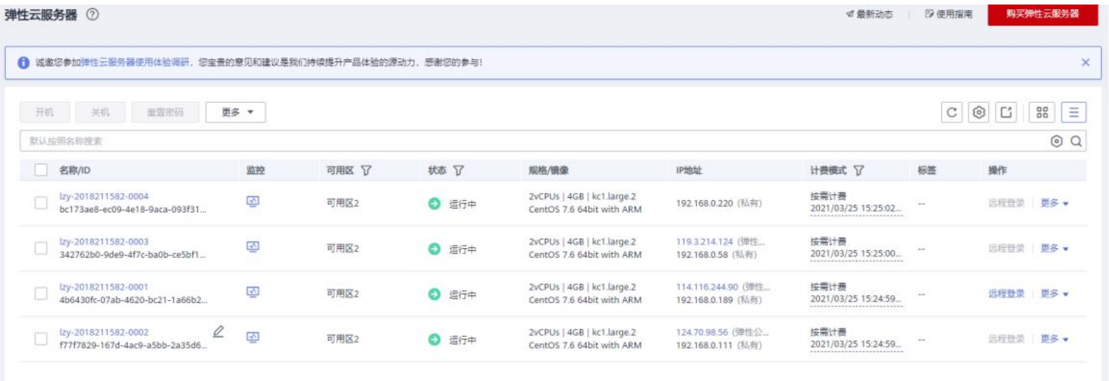
实验学生姓名：李志毅

学生班号、学号：2018211314 班 2018211582

一、实验步骤

请大家参照《云主机实现大数据实验指导书》完成本次实验，并将实验中的结果截图，完成本次实验报告。

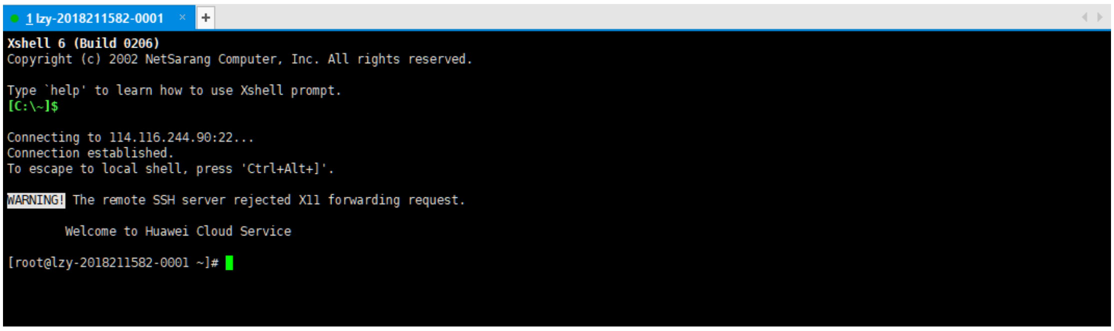
1. “1.2 购买华为云 ECS”中的步骤 6，要求自定义云服务器名称为“姓名+学号+节点序号（1234)”。



诚邀您参加弹性云服务器使用体验调研，您宝贵的意见和建议是我们持续提升产品体验的源动力，感谢您的参与！

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
lzy-2018211582-0004 bc173ae8-ec09-4e18-9aca-093f31...		可用区2	运行中	2vCPUs | 4GB | kc1.large.2 CentOS 7.6 64bit with ARM	192.168.0.220 (私有)	按需计费 2021/03/25 15:25:02...	--	远程登录 | 更多
lzy-2018211582-0003 342762b0-9de9-4f7c-ba0b-ce5bf1...		可用区2	运行中	2vCPUs | 4GB | kc1.large.2 CentOS 7.6 64bit with ARM	119.3.214.124 (弹性... 192.168.0.58 (私有)	按需计费 2021/03/25 15:25:00...	--	远程登录 | 更多
lzy-2018211582-0001 4b6430fc-07ab-4620-bc21-1a66b2...		可用区2	运行中	2vCPUs | 4GB | kc1.large.2 CentOS 7.6 64bit with ARM	114.116.244.90 (弹性... 192.168.0.189 (私有)	按需计费 2021/03/25 15:24:59...	--	远程登录 | 更多
lzy-2018211582-0002 f77f7829-167d-4ac9-a5bb-2a35d6...		可用区2	运行中	2vCPUs | 4GB | kc1.large.2 CentOS 7.6 64bit with ARM	124.70.98.56 (弹性公... 192.168.0.111 (私有)	按需计费 2021/03/25 15:24:59...	--	远程登录 | 更多

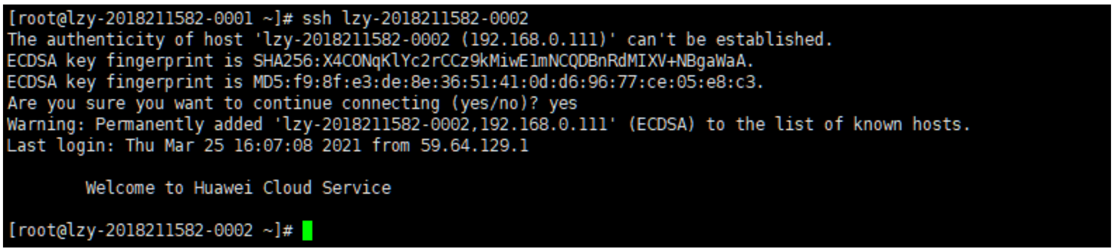
2. “2.2.1 配置 ECS”中的步骤 1，使用 putty 连接成功后，在这里贴 node1 主机登录成功的命令行界面，必须要体现出主机名和 IP 地址



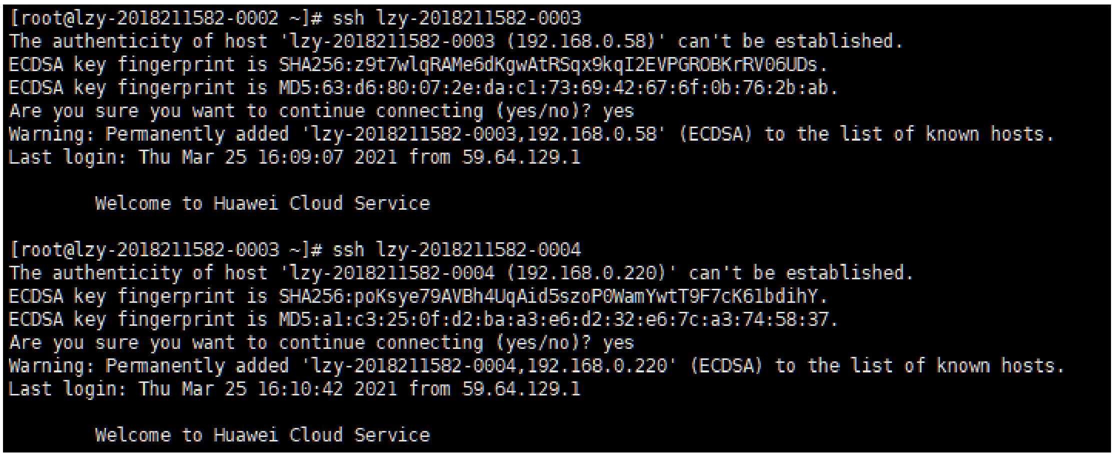
```
Xshell 6 (Build 0206)
Copyright (c) 2002 NetSarang Computer, Inc. All rights reserved.

Type `help' to learn how to use Xshell prompt.
[C:\~]$

Connecting to 114.116.244.90:22...
Connection established.
To escape to local shell, press 'Ctrl+Alt+]'.

WARNING! The remote SSH server rejected X11 forwarding request.

        Welcome to Huawei Cloud Service

[root@lzy-2018211582-0001 ~]# 
```

个人理解：我这里选用的是 xshell 登录我的 node1 服务器，使用 xshell 可以和 xftp 搭配能够上传文件。我的 node1 服务器名字为‘lzy-2018211582-0001’，图示为登录成功后显示内容，公网 ip 114.116.244.90

3. “2.2.1 配置 ECS”中的步骤 6 配置节点互信，在这里贴任意一个节点执行 ssh 命令跳转成功的截图，要体现出执行的命令和运行结果



```
[root@lzy-2018211582-0001 ~]# ssh lzy-2018211582-0002
The authenticity of host 'lzy-2018211582-0002 (192.168.0.111)' can't be established.
ECDSA key fingerprint is SHA256:X4CONqKlYc2rCCz9kMiwE1mNCQDBnRdMIXV+NBgaWaA.
ECDSA key fingerprint is MD5:f9:8f:e3:de:8e:36:51:41:0d:d6:96:77:ce:05:e8:c3.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'lzy-2018211582-0002,192.168.0.111' (ECDSA) to the list of known hosts.
Last login: Thu Mar 25 16:07:08 2021 from 59.64.129.1

        Welcome to Huawei Cloud Service

[root@lzy-2018211582-0002 ~]# 
```

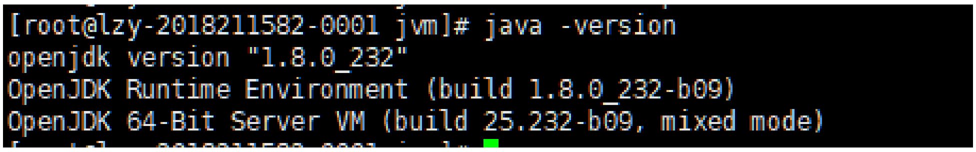


```
[root@lzy-2018211582-0002 ~]# ssh lzy-2018211582-0003
The authenticity of host 'lzy-2018211582-0003 (192.168.0.58)' can't be established.
ECDSA key fingerprint is SHA256:z9t7wlqRAMe6dKgwAtRSqx9kqI2EVPGROBKrRV06UDs.
ECDSA key fingerprint is MD5:63:d6:80:07:2e:da:c1:73:69:42:67:6f:0b:76:2b:ab.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'lzy-2018211582-0003,192.168.0.58' (ECDSA) to the list of known hosts.
Last login: Thu Mar 25 16:09:07 2021 from 59.64.129.1

        Welcome to Huawei Cloud Service

[root@lzy-2018211582-0003 ~]# ssh lzy-2018211582-0004
The authenticity of host 'lzy-2018211582-0004 (192.168.0.220)' can't be established.
ECDSA key fingerprint is SHA256:poKsye79AVBh4UqAid5szoP0WamYwtT9F7cK61bdihY.
ECDSA key fingerprint is MD5:a1:c3:25:0f:d2:ba:a3:e6:d2:32:e6:7c:a3:74:58:37.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'lzy-2018211582-0004,192.168.0.220' (ECDSA) to the list of known hosts.
Last login: Thu Mar 25 16:10:42 2021 from 59.64.129.1

        Welcome to Huawei Cloud Service
```

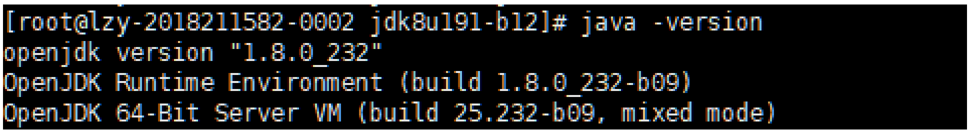
个人理解：配置 ssh 互信，使得四个节点之间可以无需密码即可 ssh 命令登录，为之后的 MapReduce 项目铺垫，这里展示了配置完成后的截图，如图所示，node1 节点 **‘lzy-2018211582-0001’** 服务器执行命令 **ssh lzy-2018211582-0002** 后可直接登录到‘**lzy-**

2018211582-0002'，在 node2 节点上执行 **ssh lzy-2018211582-0003** 可直接登陆到 node3 节点，node3 节点执行命令可直接登录到 node4 节点。

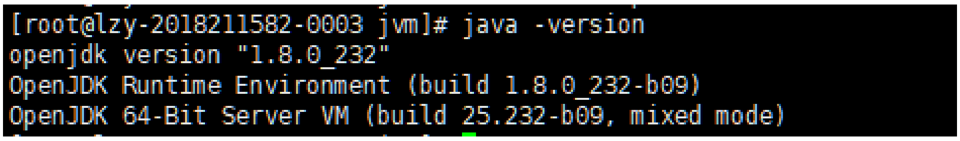
4. “2.2.2 安装 JDK”中的步骤 5，在这里贴执行“java -version”命令后的结果，显示 Java 版本即安装成功



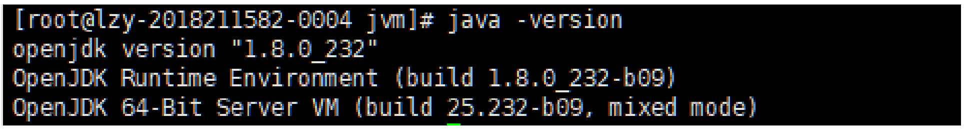
```
[root@lzy-2018211582-0001 jvm]# java -version
openjdk version "1.8.0_232"
OpenJDK Runtime Environment (build 1.8.0_232-b09)
OpenJDK 64-Bit Server VM (build 25.232-b09, mixed mode)
```



```
[root@lzy-2018211582-0002 jdk8u191-b12]# java -version
openjdk version "1.8.0_232"
OpenJDK Runtime Environment (build 1.8.0_232-b09)
OpenJDK 64-Bit Server VM (build 25.232-b09, mixed mode)
```



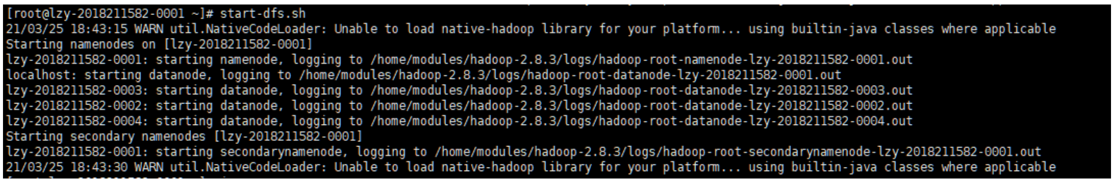
```
[root@lzy-2018211582-0003 jvm]# java -version
openjdk version "1.8.0_232"
OpenJDK Runtime Environment (build 1.8.0_232-b09)
OpenJDK 64-Bit Server VM (build 25.232-b09, mixed mode)
```



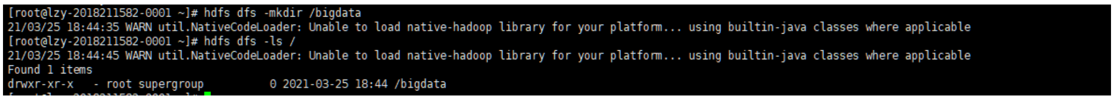
```
[root@lzy-2018211582-0004 jvm]# java -version
openjdk version "1.8.0_232"
OpenJDK Runtime Environment (build 1.8.0_232-b09)
OpenJDK 64-Bit Server VM (build 25.232-b09, mixed mode)
```

个人理解: 解压 jdk 文件，配置环境变量 java_home，执行 source 命令使其生效后，四个节点执行 **java -vesion** 后都显示 **openjdk version “1.8.0_232”**

5. “2.3.1 搭建 Hadoop 集群”中的步骤 12、13，在这里贴执行启动 hdfs 与执行 hdfs 命令的结果，要体现出执行的命令和运行结果



```
[root@lzy-2018211582-0001 ~]# start-dfs.sh
21/03/25 18:43:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [lzy-2018211582-0001]
lzy-2018211582-0001: starting namenode, logging to /home/modules/hadoop-2.8.3/logs/hadoop-root-namenode-lzy-2018211582-0001.out
localhost: starting datanode, logging to /home/modules/hadoop-2.8.3/logs/hadoop-root-datanode-lzy-2018211582-0001.out
lzy-2018211582-0003: starting datanode, logging to /home/modules/hadoop-2.8.3/logs/hadoop-root-datanode-lzy-2018211582-0003.out
lzy-2018211582-0002: starting datanode, logging to /home/modules/hadoop-2.8.3/logs/hadoop-root-datanode-lzy-2018211582-0002.out
lzy-2018211582-0004: starting datanode, logging to /home/modules/hadoop-2.8.3/logs/hadoop-root-datanode-lzy-2018211582-0004.out
Starting secondary namenodes [lzy-2018211582-0001]
lzy-2018211582-0001: starting secondarynamenode, logging to /home/modules/hadoop-2.8.3/logs/hadoop-root-secondarynamenode-lzy-2018211582-0001.out
21/03/25 18:43:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

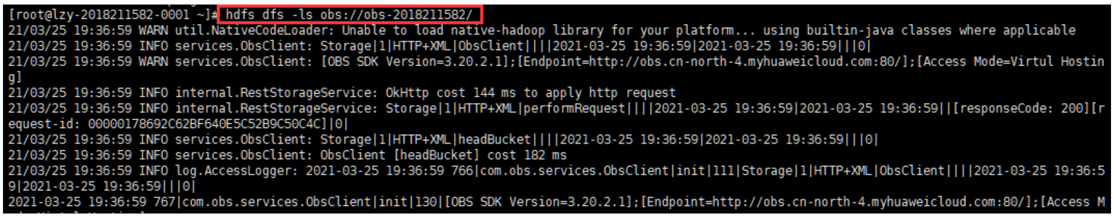


```
[root@lzy-2018211582-0001 ~]# hdfs dfs -mkdir /bigdata
21/03/25 18:44:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[root@lzy-2018211582-0001 ~]# hdfs dfs -ls /
21/03/25 18:44:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x   - root supergroup          0 2021-03-25 18:44 /bigdata
```

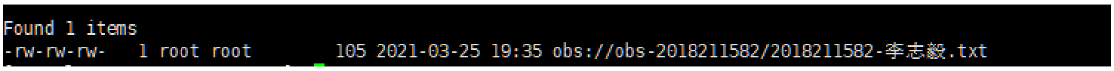
个人理解：启动 hdfs，node1 启动 namenode，node2、3、4 启动 datanode，hdfs 启动后，node1 执行 jps 可以看到 Jps、SecondaryNameNode、ResourceManager、NameNode，其余三个节点执行 jps 可以看到 NodeManager、Jps、DataNode

执行 hdfs 命令 **hdfs dfs -mkdir/bigdata**，在 hdfs 上创建 bigdata 文件夹，执行 **hdfs dfs -ls** 可以查看到该文件夹已创建成功

6. “2.3.2 测试与 OBS 互联”中，要求上传的文件以自己的学号+姓名命名，在这里贴上传成功后执行 hdfs 命令查看 OBS 文件的结果



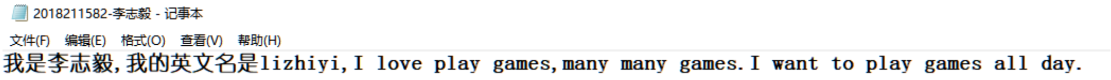
```
[root@lzy-2018211582-0001 ~]# hdfs dfs -ls obs://obs-2018211582/
21/03/25 19:36:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/03/25 19:36:59 INFO services.ObsClient: Storage|1|HTTP+XML|ObsClient||||2021-03-25 19:36:59|2021-03-25 19:36:59|||0|
21/03/25 19:36:59 WARN services.ObsClient: [OBS SDK Version=3.20.2.1];[Endpoint=http://obs.cn-north-4.myhuaweicloud.com:80/];[Access Mode=Virtul Hostin
g]
21/03/25 19:36:59 INFO internal.RestStorageService: OkHttp cost 144 ms to apply http request
21/03/25 19:36:59 INFO internal.RestStorageService: Storage|1|HTTP+XML|performRequest||||2021-03-25 19:36:59|2021-03-25 19:36:59||[responseCode: 200][r
equest-id: 00000178692C62BF640E5C52B9C50C4C]|0|
21/03/25 19:36:59 INFO services.ObsClient: Storage|1|HTTP+XML|headBucket||||2021-03-25 19:36:59|2021-03-25 19:36:59|||0|
21/03/25 19:36:59 INFO services.ObsClient: ObsClient [headBucket] cost 182 ms
21/03/25 19:36:59 INFO log.AccessLogger: 2021-03-25 19:36:59 766|com.obs.services.ObsClient|init|111|Storage|1|HTTP+XML|ObsClient||||2021-03-25 19:36:5
9|2021-03-25 19:36:59|||0|
2021-03-25 19:36:59 767|com.obs.services.ObsClient|init|130|[OBS SDK Version=3.20.2.1];[Endpoint=http://obs.cn-north-4.myhuaweicloud.com:80/];[Access M
```



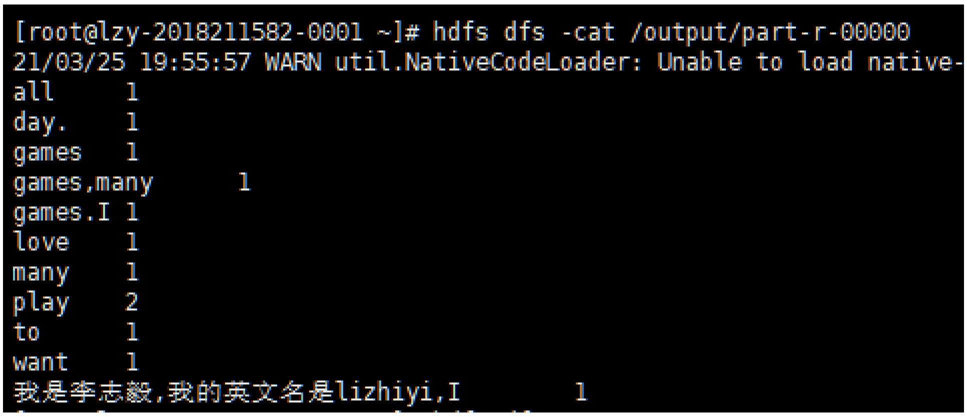
```
Found 1 items
-rw-rw-rw-   1 root root        105 2021-03-25 19:35 obs://obs-2018211582/2018211582-李志毅.txt
```

个人理解：我的 obs 桶名为“obs-2018211582”，上传文件名为“2018211582-李志毅.txt”，上传完成执行 hdfs 命令查看桶内文件可以看到该文件已经上传成功，在 OBS 中展示。

7. “3.3.1 测试 Hadoop 集群功能”中，步骤 2 的测试文件请同学们自定义文件内容，要求包含自己的姓名中英文，且至少有一个单词的数量大于等于 2，在这里贴 wordcount 的结果



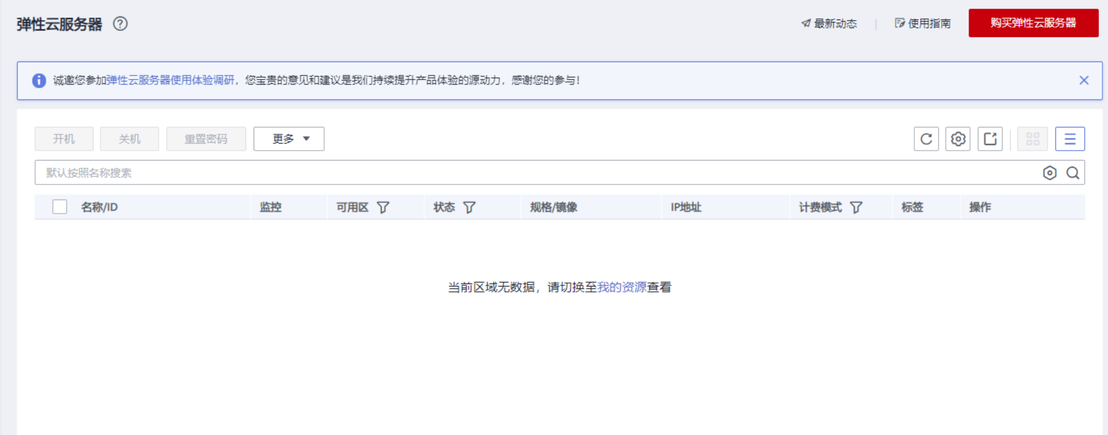
```
我是李志毅，我的英文名是lizhiyi，I love play games，many many games. I want to play games all day.
```



```
[root@lzy-2018211582-0001 ~]# hdfs dfs -cat /output/part-r-00000
21/03/25 19:55:57 WARN util.NativeCodeLoader: Unable to load native-
all     1
day.    1
games   1
games,many      1
games.I 1
love    1
many    1
play    2
to      1
want    1
我是李志毅,我的英文名是lizhiyi,I       1
```

个人理解：事先编写好 txt 文档内容，包含我的中文名和文名，以及出现两次的单词 **“play”**，执行 wordcount 命令后可以看到结果如上，其统计出 play 出现两次，其余单词一次。

8. 请同学们实验后一定按照“4 释放云服务器资源”中的说明完成 ECS 资源和 OBS 桶的释放，否则会继续计费



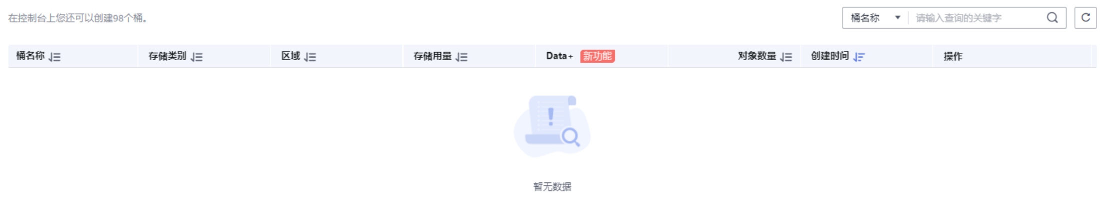
弹性云服务器 ⓘ

诚邀您参加弹性云服务器使用体验调研，您宝贵的意见和建议是我们持续提升产品体验的源动力，感谢您的参与！

默认按照名称搜索

□ 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

当前区域无数据，请切换至我的资源查看



在控制台上您还可以创建98个桶。

桶名称	存储类别	区域	存储用量	Data+ 新功能	对象数量	创建时间	操作

暂无数据

二、结果分析

1. hdfs-site.xml 中参数 dfs.replication 的含义是什么？为什么要设置为 3?

解:hdfs-site.xml 中的参数 dfs.replication 代表备份系数,即缺省的块复制数量,代指 DataNode 存储 block 的副本数量，默认是 3，此数不能大于集群的机器数，理论上 replication 值越大跑数速度越快，但是需要的存储空间也更多，默认选择 3 是因为 HDFS 采用一种称为机架感知的策略来改进数据的可靠性、可用性和网络带宽的利用率。在大多数情况下，HDFS 的副本系数是 3，HDFS 的存放策略是一个副本存放在本地机架节点上,另一个副本存放在同一机架的另一个节点上,第三个副本存放在在不同机架的节点上。这种策略减少了机架间的数据传输，提高了写操作的效率。机架错误的概率远比节点错误的概率小，所以这种策略不会对数据的可靠性和可用性造成影响。与此同时，因为数据只存在两个机架上，这种策略减少了读数据时需要的网络传输带宽。在这种策略下，副本并不是均匀地分布在机架上。这种策略在不损坏可靠性和读取性能的情况下，改善了写的性能。