

# 浙江大学

## 本科实验报告

华为云鲲鹏集群搭建

课程名称： 数据分析与算法设计

姓 名： 周灿松

学 院： 信息与工程学院

系：

专 业： 信息工程

学 号： 3190105055

指导老师： 赵明敏

2021 年 12 月 23 日

# 浙江大学实验报告

专业： 信息工程  
姓名： 周灿松  
学号： 3190105055  
日期： 2021 年 12 月 23 日  
地点： 教 7-104

课程名称： 数据分析与算法设计      指导老师： 赵明敏      成绩： \_\_\_\_\_  
实验名称： 华为云鲲鹏集群搭建      实验类型： 设计实验      同组学生姓名： \_\_\_\_\_

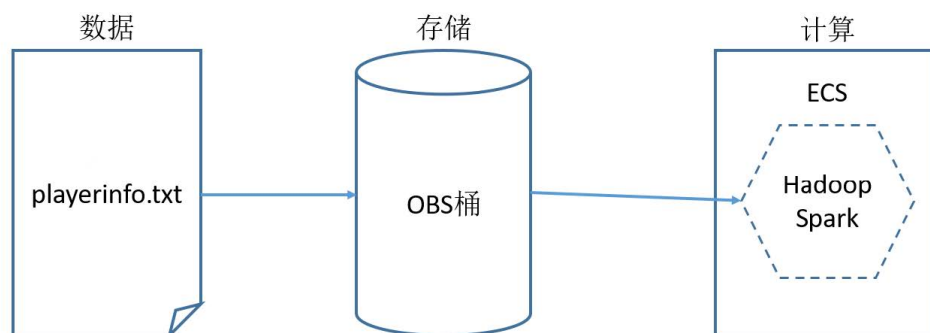
## 一 实验概览

### 1. 实验介绍

本实验基于华为云 OBS 和 华为云 ECS 服务构建一个存算分离的基本架构，并通过运行一个计算程序来完成存算分离架构的验证。本实验的实验数据存储存储在 OBS 中，通过在 ECS 上部署开源组件（Hadoop 和 Spark）构成计算环境，最后编写 Spark 程序访问存储在 OBS 上的数据进行计算（单词出现次数统计）并输出结果。

本实验的基本步骤包含：

- (1) 购买并配置 ECS；
- (2) 购买 OBS 并获取访问密钥 AK/SK 信息；
- (3) 搭建 Hadoop 集群；
- (4) 搭建 Spark 集群；
- (5) 编写 Spark 程序验证存算分离。



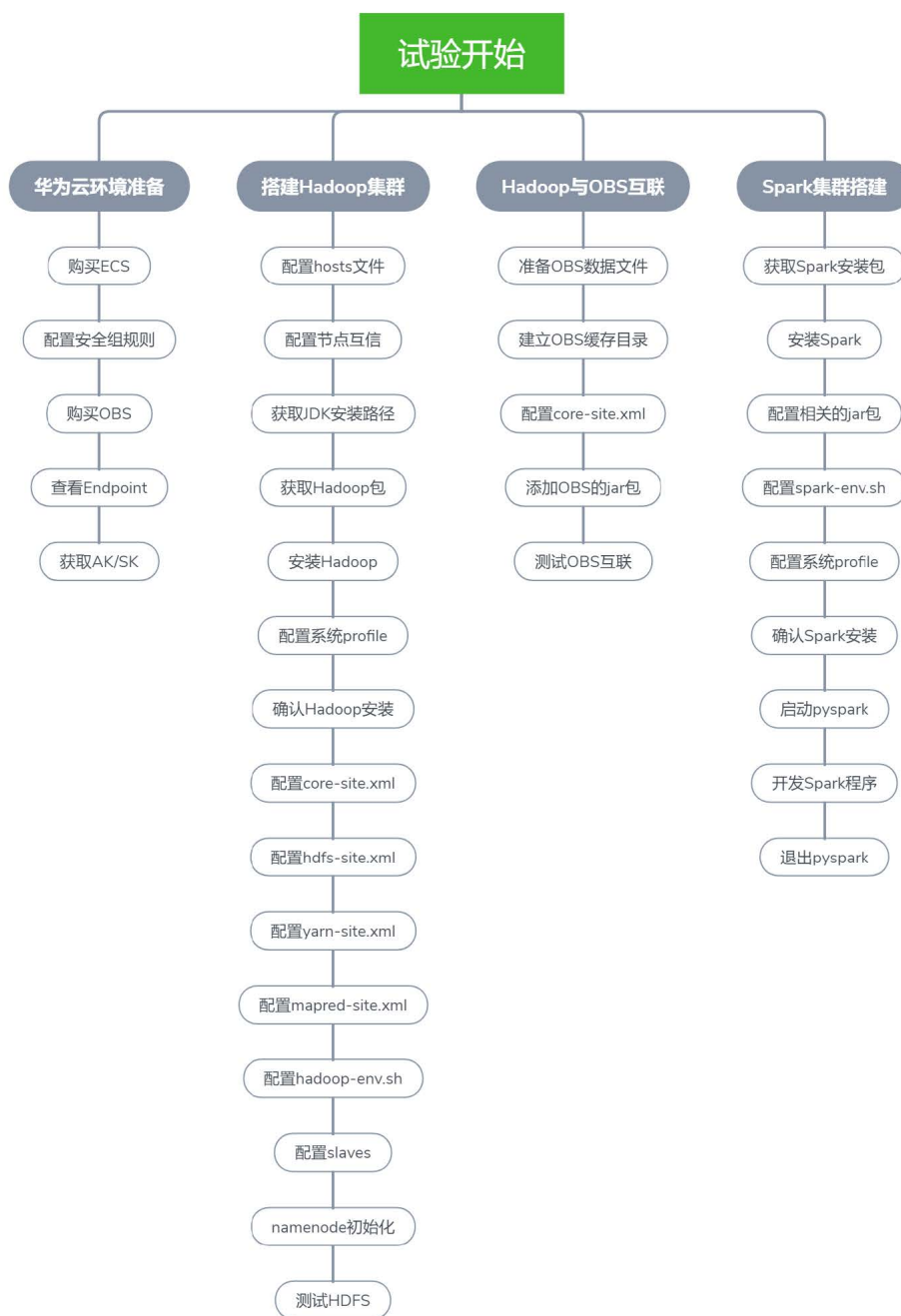
### 2. 实验目的

- (1) 掌握华为云 OBS 的购买和使用。
- (2) 掌握华为云 ECS 的购买和使用。

(3) 掌握 Hadoop  
Spark 环境搭建。

(4) 掌握 Spark 程序读取 OBS 数据。

### 3. 实验流程



## 二 华为云环境准备

按照实验手册中的方法购买华为云 ECS 服务器, 这里需要注意的一点是需要配置好安全组中的入方向规则。为了方便, 和实验手册中一样将所有协议放通并且将规则的优先级设为 1 使其成为优先级最高的规则。在完成了这一步之后便可以使用远程工具通过 ssh 连接到云服务器了。

<input type="checkbox"/>	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
<input type="checkbox"/>	ecs-zcs 96288265-0e13-46ee-8cb...		可用区2	运行中	4vCPUs   8GiB   k... openEuler 20.03 6...	121.36.27.11 (...) 192.168.0.69 (...)	按需计费 2021/12/22...	--	远程登录   更多

购买完成截图

## 三 准备 OBS 服务

### 1. 购买 OBS

按照实验教程中的步骤进行操作, 购买完成截图如下, 并记录下 endpoint 值

桶名称	存储类别	区域	数据冗余存储策...	存储用量	Data+ 新功能	对象数量	创建时间	操作
obs-zcs	--	华北-北京四	--	--		--	2021/12/22 16:...	修改存储类别   删除

桶创建完成截图

### 2. 获取访问密钥 AK/SK

按照实验教程所示操作得到访问密钥, 为后续操作做准备

## 四 搭建 Hadoop 集群

### 1. 实验介绍

#### 1.1 关于本实验

本部分实验需要在已经购买的 ECS 上搭建 Hadoop 集群, 并且通过配置与华为云 OBS 服务互联, 使 Hadoop 集群可读取 OBS 数据。

#### 1.2 实验目的

- (1) 掌握在 ECS 上搭建 Hadoop 集群方法
- (2) 掌握 Hadoop 集群与华为云 OBS 互联方法

### 2. Hadoop 集群搭建

#### 2.1 配置 ECS

此处我选择了使用自己比较熟悉的 Windows Terminal 进行远程连接, 方法为在 Windows Terminal 中输入 `ssh root@121.36.27.11`, 并且输入设置好的密码即可使用 root 用户登录云端服务器。如图

## 1所示

```
Zhou_@skyline ~
> ssh root@121.36.27.11
The authenticity of host '121.36.27.11 (121.36.27.11)' can't be established.
ECDSA key fingerprint is SHA256:0WbqXRA+6Kgs5RNB9P/ghKUwH0RmrzF3lMU+brxGZsw.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '121.36.27.11' (ECDSA) to the list of known hosts.

Authorized users only. All activities may be monitored and reported.
root@121.36.27.11's password:

Welcome to Huawei Cloud Service

Welcome to 4.19.90-2003.4.0.0036.oe1.aarch64

System information as of time: Wed Dec 22 17:24:42 CST 2021

System load:      0.19
Processes:        131
Memory used:      4.8%
Swap used:        0.0%
Usage On:         9%
IP address:       192.168.0.69
Users online:     1

[root@ecs-e492 ~]#
```

图 1: 远程登陆

随后如教程中所示在 terminal 中通过 ssh 公钥建立互信节点

## 2.2 获取 JDK 的安装路径

按照实验手册中的操作步骤得到 JAVA\_HOME 的值为

```
/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64
```

## 3. 搭建 Hadoop 伪分布式集群

### 3.1 Hadoop 安装

在安装 Hadoop 时, 实验教程中是直接让云服务器使用 wget 工具下载 Hadoop 的压缩包, 但是在这一步时, 我发现华为云的网速慢得离谱, 仅有 12kB/s. 下载一个 244MB 的文件竟然需要 15h, 这是难以容忍的。在询问相关人员无果、网上查阅资料也无法解决的情况下, 我选择了在 wsl 中下载好安装包然后上传到云服务器上的方式解决, 如图 2所示。

```
> scp hadoop-2.8.3.tar.gz root@121.36.27.11:/root

Authorized users only. All activities may be monitored and reported.
root@121.36.27.11's password:
hadoop-2.8.3.tar.gz
```

图 2: 上传 hadoop 安装包

将该压缩包解压, 并将对应文件移动到正确位置后, 使用 vim 编辑器修改系统配置文件, 修改内容如图 3 所示。

```
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64
export HADOOP_HOME=/home/modules/hadoop-2.8.3
export PATH=$JAVA_HOME/bin:$PATH
export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH
export HADOOP_CLASSPATH=/home/modules/hadoop-2.8.3/share/hadoop/tools/lib/*:$HADOOP_CLASSPATH
```

图 3: 配置系统环境变量

修改好环境变量后进入 root 目录下验证 hadoop 安装信息, 结果如图 4 所示, 正确显示了 hadoop 的版本信息, 安装成功。

```
[root@ecs-zcs ~]# hadoop version
Hadoop 2.8.3
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r b3fe56402d908019d99af1f1f4fc65cb1d1436a
Compiled by jdu on 2017-12-05T03:43Z
Compiled with protoc 2.5.0
From source with checksum 9ff4856d824e983fa510d3f843e3f19d
This command was run using /home/modules/hadoop-2.8.3/share/hadoop/common/hadoop-common-2.8.3.jar
```

图 4: 验证 hadoop 安装

### 3.2 伪分布式配置

配置伪分布式时按照华为官方实验教程进行如图 5 所示步骤

```
[root@ecs-zcs ~]# vim /home/modules/hadoop-2.8.3/etc/hadoop/core-site.xml
[root@ecs-zcs ~]# vim /home/modules/hadoop-2.8.3/etc/hadoop/core-site.xml
[root@ecs-zcs ~]# vim /home/modules/hadoop-2.8.3/etc/hadoop/hdfs-site.xml
[root@ecs-zcs ~]# vim /home/modules/hadoop-2.8.3/etc/hadoop/yarn-site.xml
[root@ecs-zcs ~]# vim /home/modules/hadoop-2.8.3/etc/hadoop/hdfs-site.xml
[root@ecs-zcs ~]# cd /home/modules/hadoop-2.8.3/etc/hadoop/
[root@ecs-zcs hadoop]# mv mapred-site.xml.template mapred-site.xml
[root@ecs-zcs hadoop]# vim /home/modules/hadoop-2.8.3/etc/hadoop/mapred-site.xml
[root@ecs-zcs hadoop]# [root@ecs-zcs hadoop]# vim /home/modules/hadoop-2.8.3/etc/hadoop/hadoop-env.sh
[root@ecs-zcs hadoop]# ^C
[root@ecs-zcs hadoop]# vim /home/modules/hadoop-2.8.3/etc/hadoop/slaves
```

图 5: 伪分布式配置

在 vim 中修改了对应的文件之后, 即可使用 JPS 查看运行的进程, 查询结果如图 6 所示

```
[root@ecs-zcs hadoop]# jps
2630 NameNode
2791 DataNode
2984 SecondaryNameNode
3114 Jps
1644 WrapperSimpleApp
```

图 6: 使用 JPS 查看启用的进程

### 3.3 Hadoop 与 OBS 互联

在上传文件时, 我选择了上传自己在另一门课中所编写的 Verilog 代码的顶层文件, 并利用这次的实验原理查询在文件中声明了多少次 wire 类型的变量 (即统计关键词 wire 的个数)。

实验过程与华为提供的实验手册操作类似, 下载添加 jar 包的步骤如图 7 所示。

```
[root@ecs-zcs ~]# cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/common/lib/
[root@ecs-zcs ~]# cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/tools/lib/
[root@ecs-zcs ~]# cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/
[root@ecs-zcs ~]# cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/hdfs/lib/
```

图 7: 添加 OBSFileSystem 相关 jar 包

随后测试 OBS 互联, 执行 HDFS 命令查看 OBS 文件, 结果如图 8 所示: 成功查询到了 OBS 桶中的文件, 互联成功。

```
Found 1 items
-rw-rw-rw-  1 root root      6424 2021-12-22 11:15 obs://obs-zcs/Risc5CPU.v
```

图 8: 测试 OBS 互联

## 五 Spark 集群搭建

### 1. 实验介绍

#### 1.1 关于本实验

本部分实验介绍安装 Spark 集群, 并使 Spark 能够读取 OBS 数据, 使用 Python 编写 Spark 程序处理 OBS 中的数据 (单词统计)。该实验使用 Spark 集群 + OBS 实现存算分离, 提高计算性能。

#### 1.2 实验目的

- (1) 掌握 Spark 集群搭建
- (2) 掌握 Spark 集群与 OBS 互联
- (3) 使用 Python 编写 Spark 程序



## 2. Spark 集群存算分离

## 2.1 搭建 Spark 集群

仍旧采用安装 `hadoop` 的方法下载好 `Spark` 的压缩包，安装实验教程中的方法将其解压、配置相关的 `jar` 包，然后配置好 `Spark` 的配置文件与系统环境变量，最后是系统环境变量生效。如图 9 所示，`Spark` 安装成功。

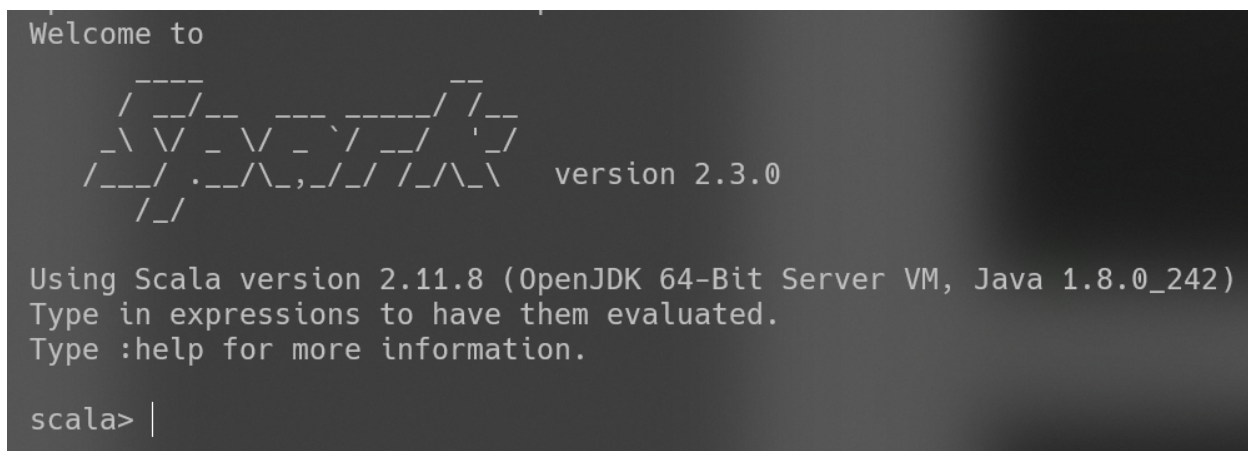


图 9: Spark 安装成功

## 2.2 验证存算分离

shell 编写如 Listing 1所示 Python 代码,即可达到查询出 wire 个数的目的,查询结果如图10所示。

Listing 1: 验证代码

```
1 # -*- coding:utf-8 -*-
2 from pyspark.sql.session import SparkSession
3 spark = SparkSession.builder.getOrCreate()
4 spark.sparkContext.setLogLevel("WARN")
5 # 读取OBS数据
6 lines = spark.read.text("obs://obs-bigdataprobe/").rdd.map(lambda r: ...
    r[0])
7 # 统计单词出现次数
8 counts = lines.flatMap(lambda x: x.split(' ')).map(lambda x: (x, ...
    1)).reduceByKey(lambda x, y: x + y)
9 output = counts.collect()
10 # 输出统计结果
11 for (word, count) in output:
12     if word == 'wire':
13         print("%s: %i" % (word, count))
```



```
>>> for (word, count) in output:
...     if word == 'wire':
...         print("%s: %i" % (word, count))
...
wire: 33
```

图 10: 查询结果

## 六 释放华为云服务

如教程所示删除 ECS 服务器，注意释放弹性公网 IP 地址。然后删除 OBS 桶中的对象之后删除 OBS 桶。

至此，本次实验结束。

## 七 心得体会

这一实验主要是利用华为的弹性云服务器以及 OBS 桶搭建了一个存算分离的 Spark 集群。在实验中，数据以对象的形式存储在 OBS 桶之中，我们通过编写的 Python 代码使用弹性云服务器访问 OBS 桶中的数据并弹性云服务器作为算力进行数据分析。

示例中的应用为统计给定文件中各个姓名出现的次数，在了解了整个 Spark 集群的工作原理之后，我对 Python 代码进行了小幅度的修改，使其功能变成了统计给定文本中特定单词出现的次数。并利用这一功能统计了自己另一个作业中变量声明的次数。虽然目前看来没有什么大的用处，但是可以预见，当需要分析的数据量变大之后，这一功能的作用也会变得更加的明显。