浙江大学

本科实验报告

华为云鲲鹏 BigData Pro 集群搭建

课程名称：	数据分析与算法设计
姓　　名：	姚桂涛
学　　院：	信息与电子工程学院
专　　业：	信息工程
学　　号：	**3190105597**
指导老师：	赵明敏

2022 年 1 月 9 日

浙江大学实验报告

专业：信息工程
姓名：姚桂涛
学号：3190105597
日期：2022 年 1 月 9 日
地点：-

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 BigData Pro 集群搭建　实验类型：-　同组学生姓名：

一、 实验介绍与目的

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。

实验的实验数据存储在 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。

计算验证的例子是统计数据中单词的出现频次，并输出结果。

二、 华为云环境准备

1. 购买华为云 ECS

弹性云服务器 ECS（Elastic Cloud Server）是一种可随时自助获取、可弹性伸缩的云服务器，帮助用户打造可靠、安全、灵活、高效的应用环境，确保服务持久稳定运行，提升运维效率。

选择云服务器 ECS，可以轻松构建具有以下优势的计算资源：

- 无需自建机房，无需采购以及配置硬件设施。
- 分钟级交付，快速部署，缩短应用上线周期。
- 快速接入部署在全球范围内的数据中心和边界网关协议 BGP（Border Gateway Protocol）机房。
- 成本透明，按需使用，支持根据业务波动随时扩展和释放资源。
- 提供 GPU 和 FPGA 等异构计算服务器、弹性裸金属服务器以及通用的 x86 架构服务器。
- 支持通过内网访问其他阿里云服务，形成丰富的行业解决方案，降低公网流量成本。
- 提供虚拟防火墙、角色权限控制、内网隔离、防病毒攻击及流量监控等多重安全方案。
- 提供性能监控框架和主动运维体系。
- 提供行业通用标准 API，提高易用性和适用性。

主要流程为：购买 ECS，配置安全组规则。
做好这些之后就可以通过远程工具进行连接了。

2. 购买华为云 OBS

对象存储服务 OBS（Object Storage Service，OBS）可以提供海量、安全、高可靠、低成本的数据存储能力，可供用户存储任意类型和大小的数据。适合企业备份/归档、视频点播、视频监控等多种数据存储场景。

可以应用于大数据分析，OBS 提供的大数据解决方案主要面向海量数据存储分析、历史数据明细查询、海量行为日志分析和公共事务分析统计等场景，向用户提供低成本、高性能、不断业务、无需扩容的解决方案。

主要流程为：购买 OBS，记录 IPv4 地址。

3. 掌握 AK/SK 的获取

访问密钥（Access Key ID/Secret Access Key，简称 AK/SK）包含访问密钥 ID（AK）和秘密访问密钥（SK）两部分，是华为云的长期身份凭证，可以通过访问密钥对华为云 API 的请求进行签名。

华为云通过 AK 识别访问用户的身份，通过 SK 对请求数据进行签名验证，用于确保请求的机密性、完整性和请求者身份的正确性。

掌握 AK/SK 后，做好记录，在后面配置 OBS 调用接口的时候，需要使用 AK/SK 进行签名验证。

三、 Hadoop 集群搭建

Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力进行高速运算和存储。Hadoop 实现了一个分布式文件系统（Distributed File System），其中一个组件是 HDFS（Hadoop Distributed File System）。HDFS 有高容错性的特点，并且设计用来部署在低廉的（low-cost）硬件上；而且它提供高吞吐量（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS 放宽了（relax）POSIX 的要求，可以以流的形式访问（streaming access）文件系统中的数据。

主要流程为：通过 PuTTY 远程登录并配置 ECS，获取 JDK 的安装路径，搭建 Hadoop 伪分布式集群，将 Hadoop 与 OBS 互联。

四、 Spark 集群搭建与验证存算分离

Spark 是使用 scala 实现的基于内存计算的大数据开源集群计算环境. 提供了 java,scala, python,R 等语言的调用接口。

本实验中使用 Spark 读取 OBS 数据，并使用 Python 编写 Spark 程序。

搭建好 Spark 后，为了支持 Python 语言，可以安装 Pyspark 工具，然后就可以编写 pyhton 程序，对 OBS 中上传的数据进行处理了。

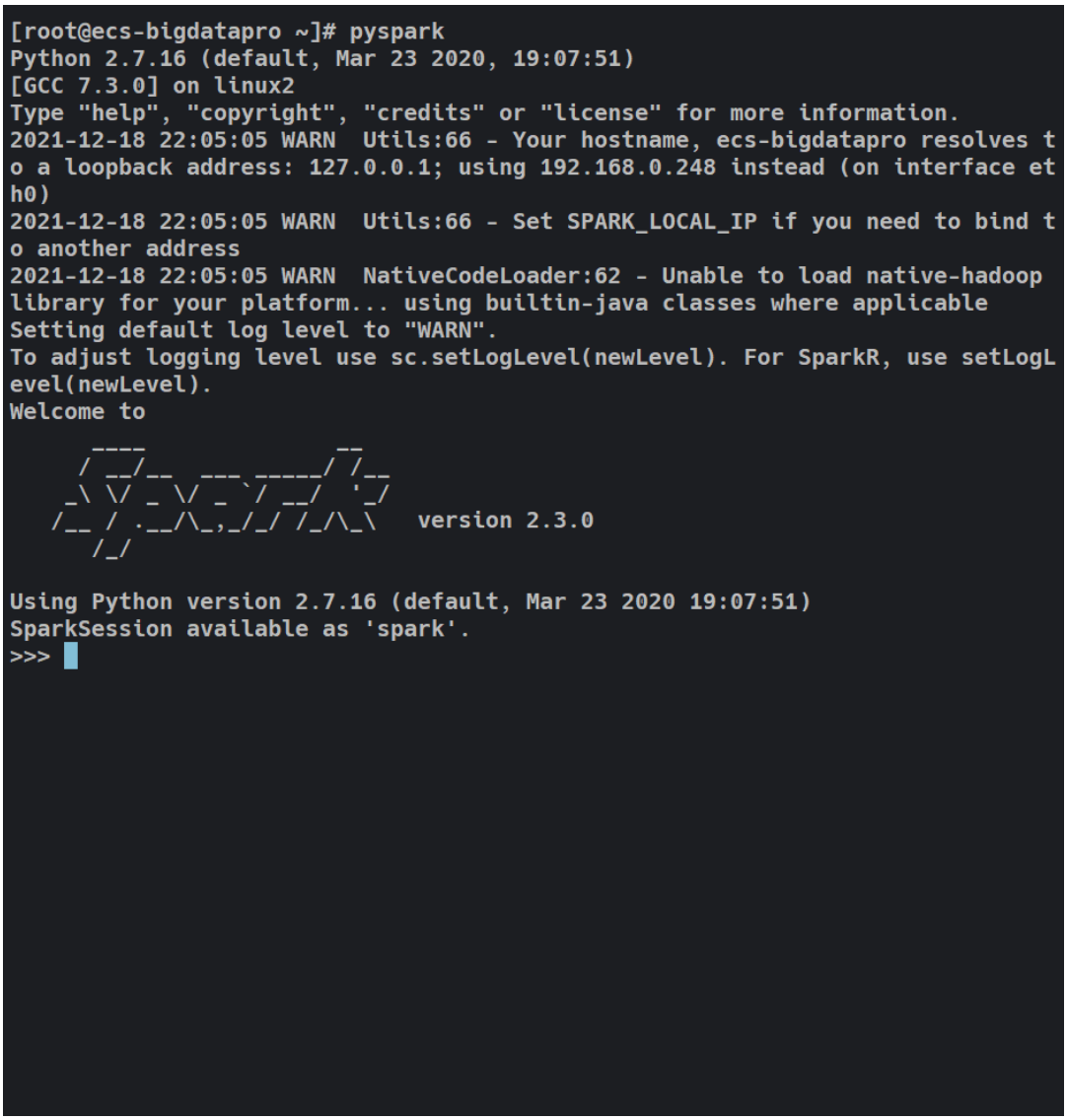

图 1: 启动 pyspark

```
>>> lines = spark.read.text("obs://oddyti-bigdatapro/").rdd.map(lambda r: r[0
])
2021-12-18 22:07:26 WARN  ObsClient:? - [OBS SDK Version=3.20.2.1];[Endpoint=
http://obs.cn-north-4.myhuaweicloud.com:5080/];[Access Mode=Virtul Hosting]
2021-12-18 22:07:26 WARN  RestStorageService:? - com.obs.services.internal.Se
rviceException: Request Error. HEAD 'http://oddyti-bigdatapro.obs.cn-north-4.
myhuaweicloud.com:5080/_spark_metadata' on Host 'oddyti-bigdatapro.obs.cn-nor
th-4.myhuaweicloud.com:5080' @ 'Sat, 18 Dec 2021 14:07:26 GMT' -- ResponseCod
e: 404, ResponseStatus: Not Found, RequestId: 0000017DCDDE710A680FE88E282BA13
4, HostId: 32AAAQAAEAABAAAQAAEAABAAAQAAEAABCSxmcjrSivKm3sjX6i8bHegNKcF4TZkE
2021-12-18 22:07:26 WARN  ObsClient:? - Storage|1|HTTP+XML|getObjectMetadata|
|||2021-12-18 22:07:26|2021-12-18 22:07:26|||404|
2021-12-18 22:07:26 WARN  RestStorageService:? - com.obs.services.internal.Se
rviceException: Request Error. HEAD 'http://oddyti-bigdatapro.obs.cn-north-4.
myhuaweicloud.com:5080/_spark_metadata%2F' on Host 'oddyti-bigdatapro.obs.cn-
north-4.myhuaweicloud.com:5080' @ 'Sat, 18 Dec 2021 14:07:26 GMT' -- Response
Code: 404, ResponseStatus: Not Found, RequestId: 0000017DCDDE7113680FE8A20875
F02F, HostId: 32AAAQAAEAABAAAQAAEAABAAAQAAEAABCSWQrRjYXV0C4Q0EFSSuKcrzTO72W5R
b
2021-12-18 22:07:26 WARN  ObsClient:? - Storage|1|HTTP+XML|getObjectMetadata|
|||2021-12-18 22:07:26|2021-12-18 22:07:26|||404|
```

图 2: 读取 OBS 桶的内容

```
>>> counts = lines.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).redu
ceByKey(lambda x, y: x + y)
>>> output = counts.collect()
[Stage 0:>                                                          (0 + 1) /

>>>
```

图 3: 存算分离代码

```
>>> for (word, count) in output:
...     print("%s: %i" % (word, count))
...
William: 1
Alex: 1
Kerry: 1
James: 2
Robert: 1
Genu: 2
Robertm: 1
Vera: 2
Olivia: 2
Edith: 2
Lax: 2
Hale: 1
Mary: 2
>>>
```

图 4: 存算分离结果

五、 释放华为云服务

做完实验后，需要释放我们购买的 ECS 与 OBS。