

## 0. 前置要求

### 平台使用前置要求

- 已经注册了华为云的开发者平台，并登记了 ID 信息，获得了华为云代金券

### 完成作业前置要求

- 安装了 Java，jdk 环境为 1.8 (Java 8)
- 了解如何构建 jar 包，如何配置 Maven 环境
- 推荐使用 IntelliJ IDEA (非必须)

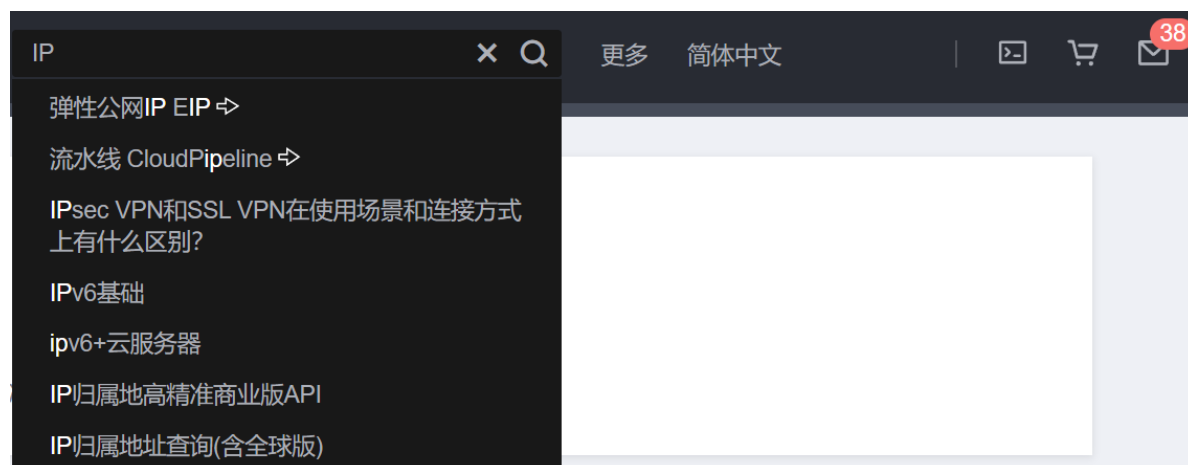
## 1. 资源购买

### 1.1 购买弹性公网 IP

登录华为云账号后，点击控制台进入控制台管理页面。



右上角搜索 IP，点击 弹性公网IP EIP。



点击 购买弹性公网IP。



选择如图所示的配置：按需计费、全动态BGP、按流量计费等。

计费模式

包年/包月

按需计费

区域

华东-上海一

不同区域的云服务产品之间内网互不相通；请就近选择靠近您业务的区域，可减少网络时延，提高访问速度。

线路

全动态BGP

静态BGP

不低于99.95%可用性保障

公网带宽

按带宽计费

流量较大或较稳定的场景

按流量计费

流量小或流量波动较大场景

加入共享带宽

多业务流量错峰分布场景

按照您实际使用的出云流量收取流量费，带宽大小仅做限速使用，不做为收费依据。  
EIP与实例解绑后，会停止收取流量费，新增IP保有费。更多计费信息请参考 [计费说明](#)

带宽大小 (Mbit/s)

5

10

20

50

100

自定义

—

20

+

带宽范围：1-300 Mbit/s

提供基础DDoS防护能力。 [了解更多](#)

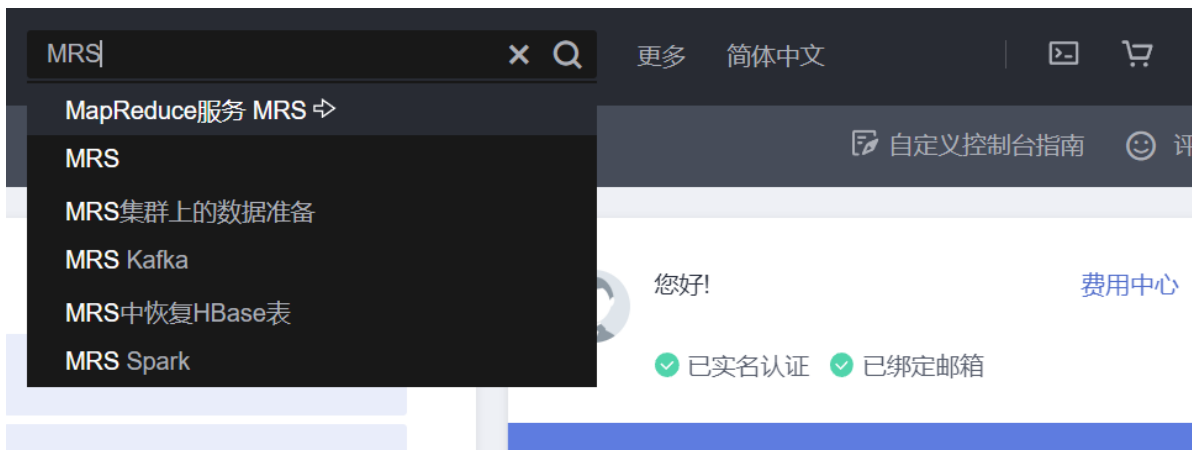
IPv6转换

☐ 一键开启，实现对外提供IPv6访问能力

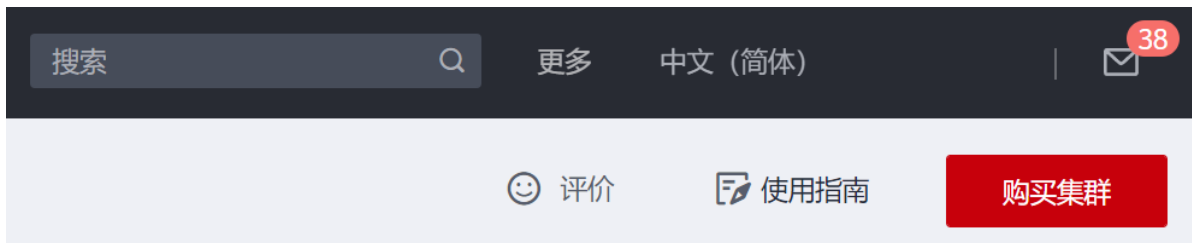
公测期间IPv6转换功能免费。

## 1.2 购买 MRS 集群

右上角搜索 MRS，点击 `MapReduce服务 MRS`。



点击 `购买集群`。



1.2.1 软件配置

选择如下图所示的配置，集群名称请填写自己的学号和姓名首字母。

MapReduce服务

快速购买

自定义购买

1 软件配置

2 硬件配置

3 高级配置

区域

华东-上海一

不同区域的资源之间内网不互通。请选择靠近您客户的区域，可以降低网络

集群名称

mrs\_flink

?

版本类型

普通版

LTS版

?

集群版本

MRS 1.9.2

集群类型

分析集群

流式集群

混合集群

必选组件默认勾选，被依赖的组件会被自动勾选。

必选组件默认勾选，被依赖的组件会被自动勾选。

分析组件

<input checked="" type="checkbox"/>	组件名	版本
<input type="checkbox"/>	Presto	0.216
<input checked="" type="checkbox"/>	Hadoop	2.8.3
<input checked="" type="checkbox"/>	Spark	2.2.2
<input type="checkbox"/>	HBase	1.3.1
<input type="checkbox"/>	Opentsdb	2.3.0
<input checked="" type="checkbox"/>	Hive	2.3.3
<input type="checkbox"/>	Hue	3.11.0
<input type="checkbox"/>	Loader	2.0.0
<input checked="" type="checkbox"/>	Tez	0.9.1
<input checked="" type="checkbox"/>	Flink	1.7.0
<input type="checkbox"/>	Alluxio	2.0.1
<input type="checkbox"/>	Ranger	1.0.1

元数据

本地元数据

数据连接

?

1.2.2 硬件配置

选择如下图所示的配置。

1 软件配置

2 硬件配置

3 高级配置

计费模式

包年/包月

按需计费

可用区

IES-cnbox

IES-liqoc

IES-ypcjy

IES-zsqvj

可用区1

可用区2

可用区3

虚拟私有云

vpc-default

查看虚拟私有云

子网

subnet-default(192.168.0.0...

查看子网

当前子网剩余可用IP数: 241

安全组

自动创建

管理安全组

弹性公网IP

管理弹性公网IP

CPU架构

x86计算

鲲鹏计算

集群节点

节点类型 ?	计费模式	实例规格	实例数量
Master节点 ?	按需计费	鲲鹏通用计算增强型 4 vCPUs   16 GB   kc1.xlarge.4 系统盘 高IO 100 GB x 1 数据盘 高IO 200 GB x 1	2
分析Core节点 ?	按需计费	鲲鹏通用计算增强型 4 vCPUs   16 GB   kc1.xlarge.4 系统盘 高IO 100 GB x 1 数据盘 高IO 100 GB x 1	<div><div>-</div><div>2</div><div>+</div></div>
分析Task节点 ?	按需计费		

1.2.3 高级配置

标签、主机名前缀、弹性伸缩、引导操作 维持默认。

标签

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中创建预定义标签。[查看预定义标签](#)

标签键

标签值

你还可以添加10个标签。

主机名前缀

用作集群中ECS机器或BMS机器主机名的前缀

弹性伸缩

请先返回上一步设置Task节点规格后再设置弹性伸缩策略。

引导操作

名称	执行节点	执行时机	操作
<div>添加</div>			

引导操作添加的脚本个数不能超过18个。

关闭 告警。

委托

暂不绑定

MRS\_ECS\_DEFAULT\_AGENCY

现有委托

数据盘加密

关闭

开启

告警

关闭

开启

集群运行异常或系统故障时，维护人员可根据告警信息定位问题原因，建议开启。

日志记录

关闭

开启

kerberos 认证 选择关闭，设置 MRS 平台的登录密码。

Kerberos认证

用户名

admin

密码

.....

该密码用于登录集群管理页面。

确认密码

.....

设置服务器 root 用户的登录密码。

登录方式

密码

密钥对

用户名

root

密码

.....

该密码用于远程登录ECS机器或BMS机器。

确认密码

.....

点击 **确认授权**，开通所有的访问规则。

通信安全授权

☒ 确认授权

授权MRS集群创建和切换子网时开通相应的访问控制规则，从而使得用户可以通过MRS管理控制台进行大数据组件部署和后续集群的使用、运维和管理等操作，此时不授权将无法创建集群。[了解更多](#)

需要开通的访问控制规则

协议端口	类型	源地址	描述
TCP : 9022	IPv4	100.125.2.76/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.2.75/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.64.105/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.64.106/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.2.79/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.2.78/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.2.77/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.65.56/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.65.54/32	MRS 默认访问控制规则
TCP : 9022	IPv4	100.125.65.55/32	MRS 默认访问控制规则

购买成功后，系统将会自动创建集群，请耐心等待集群部署完成。

1.4 检查集群信息

集群配置成功后，概览信息如下图所示。

<

mrs\_flink

使用指南

评价

下载认证凭证

概览

节点管理

组件管理

告警管理

文件管理

作业管理

租户管理

备份恢复

引导操作

弹性伸缩

标签

基本信息

了解更多

集群名称

mrs\_flink

集群状态

运行中

集群管理页面

前往 Manager

付费类型

按需计费

集群版本

MRS 1.9.2

集群类型

分析集群

IAM用户同步

未同步

同步

集群ID

a0f203b6-41ac-4a60-a371-50c78c6b2b36

创建时间

2022/10/12 10:39:53 GMT+08:00

可用区

可用区1

虚拟私有云

vpc-default

默认生效子网

subnet-default

切换子网

数据连接

单击管理

委托

-- 管理委托

弹性公网IP

添加安全组规则

Kerberos认证

关闭

日志记录

关闭

安全组

mrs\_mrs\_flink\_fxBC

通信安全授权

开启

## 2. 配置相关资源

### 2.1 建立 OBS 并行文件系统

如前文所叙步骤，登录控制台，搜索 OBS，进入对象存储服务控制台。点击 **并行文件系统** > **创建并行文件系统**。

名称任意填写，并按如图所示配置创建一个并行文件系统。

复制并行文件系统配置

选择源并行文件系统

该项可选。选择后可复制源并行文件系统的以下配置：区域 / 数据冗余策略 / 策略 / 默认加密 / 归档数据直读 / 企业项目 / 标签。

区域

华东-上海一

- 不同区域的资源之间内网互不相通，请选择靠近您业务的区域，可以降低网络时延，提高访问速度。并行文件系统创建成功后不支持变更区域，请谨慎选择。

- 目前并行文件系统不支持专属云场景。[如何选择区域](#)

文件系统名称

mrs-flink-obs

① 不能和本用户已有的文件系统重名

① 不能和其他用户已有的文件系统重名

① 创建成功后不支持修改

数据冗余存储策略

多AZ存储

单AZ存储

②

① 启用后不支持修改。多AZ存储采用相对较高计费标准。[价格详情](#)

数据在同区域的多个AZ中存储，可用性更高。

采用单AZ创建的文件系统，数据将只存储在一个可用区内，适用于对访问时延要求较低的数据存储。

策略

私有

公共读

公共读写

复制策略

②

其他用户在未经授权的情况下均无访问权限。

归档数据直读

开启

关闭

②

关闭归档直读，归档存储类别的数据要先恢复才能访问。归档存储数据恢复和访问会收取相应的费用。[价格详情](#)

### 2.2 创建委托策略

控制台处搜索 **统一身份认证服务**，进入管理页面后，点击左侧 **委托**。

统一身份认证服务

用户

用户组

权限管理

项目

委托

身份提供商

安全设置

用户

IAM用户登录链接 <https://auth.huaweicloud.com/authui/login?id=>

删除

编辑

您还可以创建49个用户。

<input type="checkbox"/>	用户名	描述	状态
<input type="checkbox"/>	企业管理员		启用

创建如下图所示委托(委托名称任意)，点击下一步。

委托 / 创建委托

★ 委托名称

mrs\_ecs\_obs

★ 委托类型

☐ 普通帐号  
将帐号内资源的操作权限委托给其他华为云帐号。

☒ 云服务  
将帐号内资源的操作权限委托给华为云服务。

★ 云服务

弹性云服务器 ECS 裸金属服务器 BMS

★ 持续时间

永久

描述

请输入委托信息。

0/255

下一步

取消

在弹出授权页面的搜索框内，搜索 `OBS OperateAccess` 策略，勾选 `OBS OperateAccess`。

单击下一步，选择权限范围方案，默认选择 `所有资源`，单击 `展开其他方案`，选择 `全局服务资源`。

在弹出的提示框中单击 `知道了`，开始授权。界面提示授权成功。单击完成，委托成功创建。

< 授权

1 选择策略

2 设置最小授权范围

3 完成

回到目录

新建策略

委托“mrs\_ecs\_obs”将拥有所选策略

查看已选(0)

从其他区域项目复制权限

全部类型

所有云服务

OBS OperateAccess

×

Q

☐

名称

类型

☐

▼

OBS OperateAccess

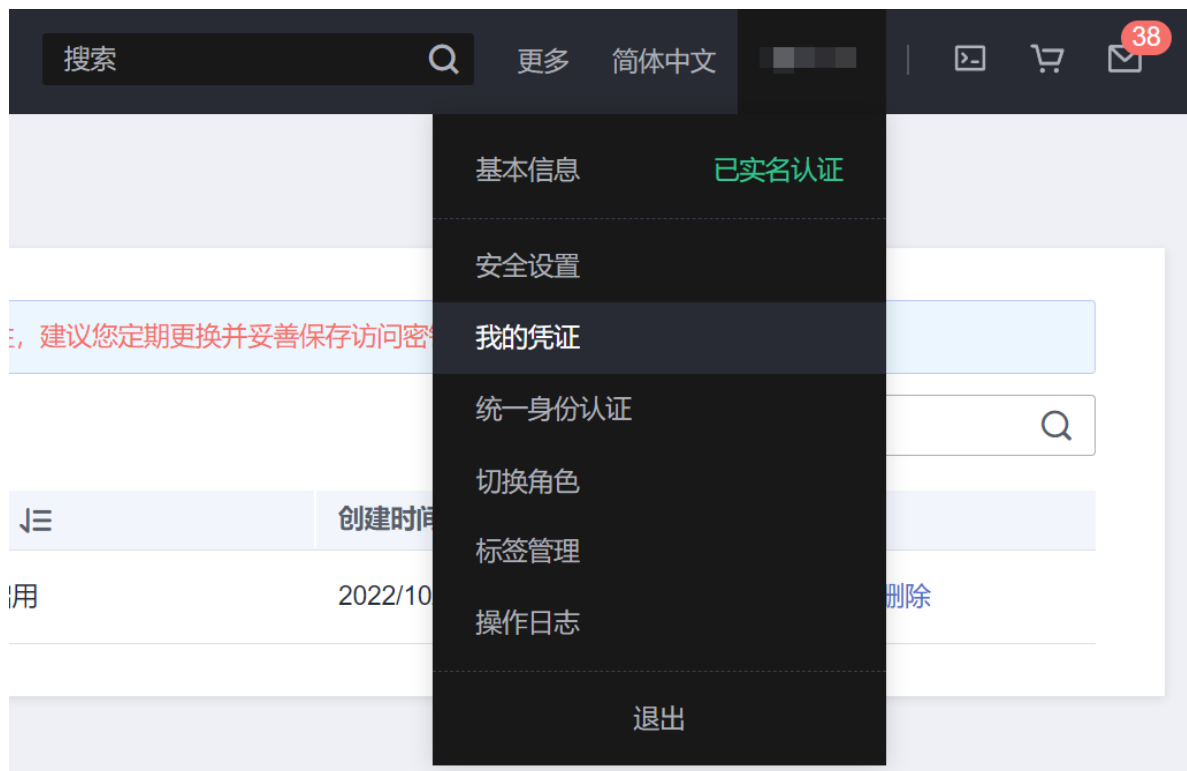
具有对象存储服务（OBS）查看桶列表、获取桶元数据、列举桶内对象、查询桶位置、上传对象、获取对象、删除对象、获取对象ACL等对象

系统策略

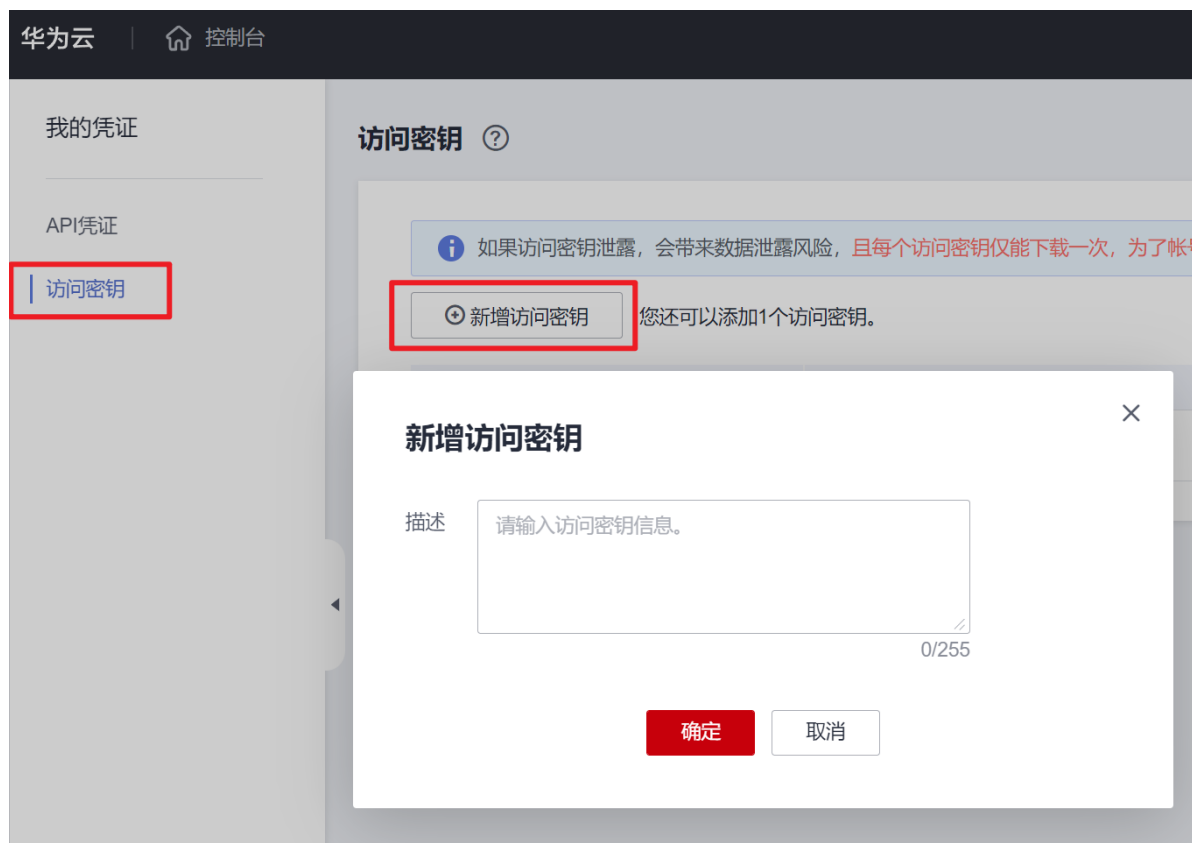


## 2.3 配置 OBS 访问密钥

点击网页右上角个人 ID，点击 **我的凭证**。



点击 **访问密钥**，点击 **+新增访问密钥**，创建一个访问密钥，并及时下载密钥信息，该密钥信息无法再次获取。下载后会得到一个名为 `credentials.csv` 的文件，里面记载了用户名，Access Key ID 和 Secret Access Key ID。



## 2.4 修改安全组策略

进入 MRS 控制台，进入节点管理页面，展开 master 节点，进入任意一个服务器节点。

< | mrs\_flink

概览 | 节点管理 | 组件管理 | 告警管理

配置Task节点

节点操作 ▼

节点组名称

^ master\_node\_default\_group

<input type="checkbox"/>	节点名称	IP
<input type="checkbox"/>	★ node-master1fcjl.mrs-zklu.com	192.168.0
<input type="checkbox"/>	★ node-master2gNfS.mrs-zklu.com	192.168.0

▽ core\_node\_analysis\_group

点击安全组，点击下方的 配置规则。

<

a0f203b6-41ac-4a60-a371...

基本信息

云硬盘

弹性网卡

安全组

弹性公网IP

192.168.0.143 (主)

全部安全组(1)

排序

更改安全组

1 mrs\_mrs\_flink\_fXBC

配置规则

点击 入方向规则，点击 一键放通，将如图所示的端口及IP源地址全部放通。

<

mrs\_mrs\_flink\_fXBC

基本信息

入方向规则

出方向规则

关联实例

添加规则

快速添加规则

删除

一键放通

通过指定属性的关键字搜索

☐

优先级 ?

☐

策略 ?

☐

协议端口 ?

1

允许

TCP : 3389

1

允许

TCP : 20-21

1

允许

ICMP : 全部

1

允许

TCP : 443

1

允许

TCP : 22

1

允许

TCP : 80

1

允许

TCP : 9022

!

一键放通功能将放通下列常用端口。

i

一键放通功能仅判断是否已添加相应的安全组规则，请确保当前安全组下没有优先级更高的拒绝策略

优先级	策略	协议端口	类型	源地址
1	允许	TCP : 22	IPv4	0.0.0.0/0 ?
1	允许	TCP : 3389	IPv4	0.0.0.0/0 ?
1	允许	TCP : 80	IPv4	0.0.0.0/0 ?
1	允许	TCP : 443	IPv4	0.0.0.0/0 ?
1	允许	TCP : 20-21	IPv4	0.0.0.0/0 ?
1	允许	ICMP : 全部	IPv4	0.0.0.0/0 ?

## 3. 运行样例程序

### 3.1 通过集群终端运行 Flink 样例程序 WordCount.jar

WordCount.jar 是 Flink 大数据处理系统的“Hello World”。它计算文本集合中单词的频率。该程序分两个步骤工作：首先，文本将文本拆分为单个单词；其次，对单词进行分组和计数。

本节将以 WordCount.jar 程序为例，说明如何通过 shell 指令创建并运行 Flink 作业。

#### 3.1.1 上传数据至 OBS 文件系统

准备一个内容为英语段落的测试文本(文本中仅含大小写字母、空格、换行符)。这里给出一段测试文本作为样例，可以将以下内容复制进 txt 文本文件中：

```
WordCount is the “Hello World” of Big Data processing systems. It computes the frequency of words in a text collection. The algorithm works in two steps: First, the texts are splits the text to individual words. Second, the words are grouped and counted.
```

进入 OBS 控制台页面，在“文件”页签下单击“新建文件夹”，分别新建 input、output 文件夹。在 input 文件夹中上传该 txt 文件。



**i** 上传操作将产生请求费用，上传成功后将产生存储费用。

✕

上传至 input/

存储类别

标准存储

上传文件

注意：并行文件系统内如有同名文件，将被新上传的文件覆盖。

清空列表

添加文件

1/100 文件 大小 258 bytes

名称 

大小 

操作

input.txt

258 bytes

删除

下一步：高级配置 (可选)

上传

取消

以 OBS 系统名为 mrs-flink 为例，该 txt 文件的最终访问路径为 `obs://mrs-flink/input/input.txt`。

## 3.2 运行样例程序

如 2.4 节第一步所示，登录任意一台 master 服务器节点，点击 `远程登录`，并以 root 账号登录 shell 终端。本节内容所提到的 OBS 文件系统中的文件名，路径等均为样例，实际操作时请自行对应修改相应的值。



登录集群客户端节点，进入客户端安装目录。输入以下指令：

```
su - omm
cd /opt/client
source bigdata_env
```

执行以下命令验证集群是否可以访问 OBS。

```
hdfs dfs -ls obs://mrs-flink/input
```

进入 Flink 样例程序目录。

```
cd /opt/client/Flink/flink/examples/batch
```

提交 Flink 作业，指定源文件数据进行消费。其中，`obs.access.key` 和 `obs.secret.key` 为 2.3 节对应的 OBS 访问密钥。

WordCount.jar 输入参数为：`--input <path> --output <path>`。

```
flink run -m yarn-cluster ./WordCount.jar -Dfs.obs.access.key=XXXX -  
Dfs.obs.secret.key=XXXX --input obs://mrs-flink/input/input.txt --output  
obs://mrs-flink/output/output2.txt
```

若作业执行成功，则显示以下输出：

```
...  
Cluster started: Yarn cluster with application id application_1654672374562_0011  
Job has been submitted with JobID a89b561de5d0298cb2ba01fbc30338bc  
Program execution finished  
Job with JobID a89b561de5d0298cb2ba01fbc30338bc has finished.  
Job Runtime: 12043 ms
```

在 OBS 文件系统中指定的结果输出文件中可查看数据分析输出的结果。下载 output.txt，查看输出结果：

```
a 1  
algorithm 1  
and 1  
are 2  
big 1  
collection 1  
  
...  
  
texts 1  
the 6  
to 1  
two 1  
wordcount 1  
words 3  
works 1  
world 1
```

### 3.3 使用华为 MRS 平台提交 WordCount.jar 程序

进入 OBS 文件系统平台，上传附件中的 WordCount.jar 程序。

进入 MRS 平台，点击作业管理，添加作业，并进行如图所示的操作（相应参数自行修改）：

← mrs\_flink

使用指南 评价 0 0

概览 节点管理 组件管理 告警管理 文件管理 作业管理 租户管理 备份恢复 引导操作

添加作业

★ 作业类型

Flink

★ 作业名称

flink\_work

★ 执行程序路径

obs://mrs-flink/input/WordCount.jar

HDFS

OBS

运行程序参数 ?

参数

值

+

执行程序参数 ?

-Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX --input obs://mrs-flink/input/input.txt --output obs://mrs-flink/output/output2.txt

HDFS

OBS

服务配置参数 ?

参数

值

+

命令参考

```
flink run -d -m yarn-cluster obs://mrs-flink/input/WordCount.jar
-Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX --input
obs://mrs-flink/input/input.txt --output obs://mrs-
flink/output/output2.txt
```

确定

取消

运行成功后，进入 OBS 平台可以看到 output/output2.txt 文件已经生成。

## 4. WordCount.jar 源码分析

代码如下：

```
package org.example;

import org.apache.flink.api.common.functions.FlatMapFunction;
import org.apache.flink.api.java.DataSet;
import org.apache.flink.api.java.ExecutionEnvironment;
import org.apache.flink.api.java.tuple.Tuple2;
import org.apache.flink.api.java.utils.ParameterTool;
import org.apache.flink.core.fs.FileSystem;
import org.apache.flink.util.Collector;

public class BatchJob {

    public static void main(String[] args) throws Exception {
        String input = null;
        String output = null;

        ParameterTool params = ParameterTool.fromArgs(args);

        try {
            input = params.getRequired("input");
```

```

        output = params.getRequired("output");
    } catch (RuntimeException e) {
        System.out.println("Argument Error");
        e.printStackTrace();

        return;
    }

    ExecutionEnvironment env =
    ExecutionEnvironment.getExecutionEnvironment();

    env.setParallelism(1);

    DataSet<String> text = env.readTextFile(input);

    DataSet<Tuple2<String, Integer>> counts = text.flatMap(new
    Tokenizer()).groupBy(0).sum(1);

    counts.writeAsText(output, FileSystem.WriteMode.OVERWRITE);

    env.execute("Flink Batch Java API Skeleton");
}

public static class Tokenizer implements FlatMapFunction<String,
    Tuple2<String, Integer>> {

    @Override
    public void flatMap(String value, Collector<Tuple2<String, Integer>>
    out) throws Exception {
        String[] tokens = value.toLowerCase().split("\\W+");

        for (String token : tokens) {
            if (token.length() > 0) {
                System.out.println(token);
                out.collect(new Tuple2<String, Integer>(token, 1));
            }
        }
    }
}
}
}

```

Flink 程序的第一步是创建一个 `StreamExecutionEnvironment`。这是一个入口类，可以用来设置参数和创建数据源以及提交任务：

```
ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();
```

为了便于演示，将结果输出于同一个文件中，设置并行度为 1：

```
env.setParallelism(1);
```



这创建了一个字符串类型的 `DataSet`。`DataSet` 是 Flink 中做数据处理的 API，上面定义了非常多的操作(如，过滤、转换、聚合、窗口、关联等)。本示例中，首先要将字符串数据解析成单词和次数(使用 `Tuple2<String, Integer>` 表示)，第一个字段是单词，第二个字段是次数，次数初始值都设置成了 1。实现了一个 `flatMap` 来做解析的工作，因为一行数据中可能有多个单词。

即：处理数据，切分(`flatMap`, `split`)，分组(`groupBy`)，统计(累加 `sum`)。

```
DataSet<String> text = env.readTextFile(input);

DataSet<Tuple2<String, Integer>> counts = text.flatMap(new
Tokenizer()).groupBy(0).sum(1);
```

`flatMap` 实现的接口 `Tokenizer()` 如下：

```
public static class Tokenizer implements FlatMapFunction<String, Tuple2<String,
Integer>> {

    @Override
    public void flatMap(String value, Collector<Tuple2<String, Integer>> out)
throws Exception {
        String[] tokens = value.toLowerCase().split("\\W+");

        for (String token : tokens) {
            if (token.length() > 0) {
                System.out.println(token);
                out.collect(new Tuple2<String, Integer>(token, 1));
            }
        }
    }
}
```