

**大数据处理实验报告**

实验四：大数据实时分析

姓 名：刘日星

学 号：X2020I1007

院 系：计算机科学与技术

专 业：计科与金融

年 级：CS1804（交换）

指导教师：石宣化

2021年 12 月 21 日

一：实验目的

1、了解大数据实时分析的用途

2、掌握大数据实时分析的基本命令

二：实验要求

1. 第四节中的实验内容要附上完整的实验过程截图以及必要的文字说明，每个人的IP地址等不同，不能直接套用样例的截图。
2. 请同学们在完成报告后，将报告的pdf版本命名为：大数据实验四+姓名+学号.pdf，并在这周五前,发到邮箱:

[aaaaltaaaa@126.com(石老师班)](mailto:aaaaltaaaa@126.com(石老师班)，)

798792873@qq.com(郑老师班)

三：实验环境配置

### 登录华为云控制台

登录华为云（网址：<https://www.huaweicloud.com>）

使用浏览器打开华为云页面，点击右上角的“登录”。

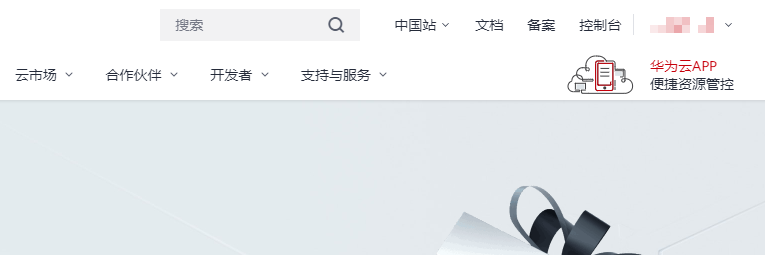


选择“账号登录”，然后输入账号、密码完成登录。

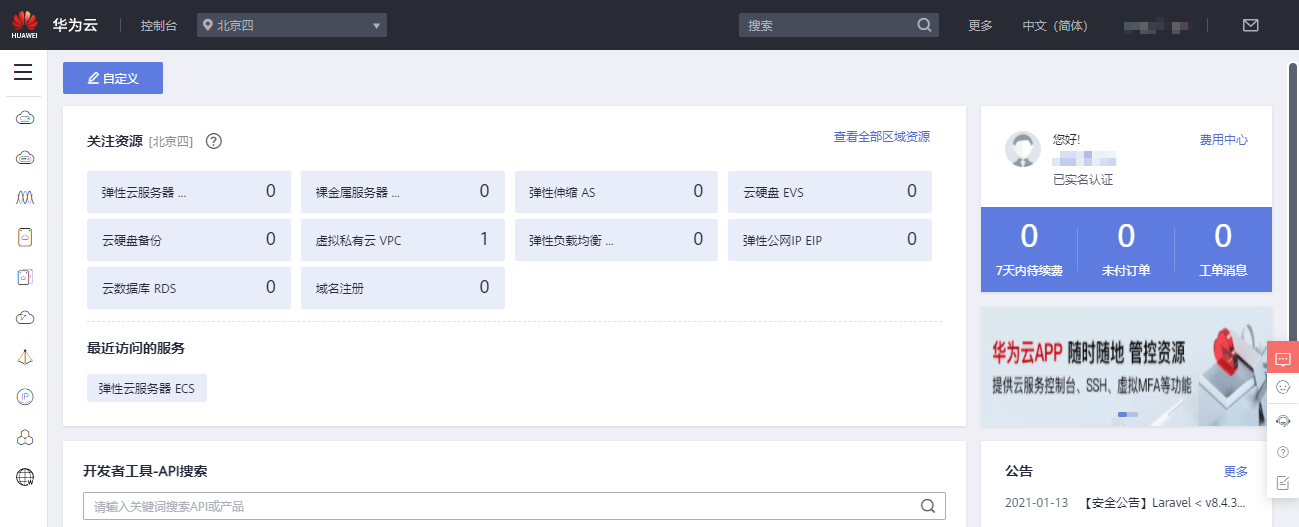


进入控制台

登录成功后点击“控制台”。



进入到华为云控制台，选择区域为“北京四”。



### 任务一：申请弹性公网IP

1. 选择弹性公网IP

进入控制台后，鼠标移动到左侧三根横线处将“服务列表”展开，并在列表中点击“网络”分类里面的“弹性公网IP EIP”。



1. 点击右上角的“购买弹性公网IP”



1. 相关配置

计费模式选择“按需计费”，区域选择“华北-北京四”，线路选择“全动态BGP”，公网带宽选择“按流量计费”，带宽大小选择“50”，购买量选择“2”。







点击右下角的“立即购买”。



1. 确认购买

查看配置详情是否正确。



确认无误后点击右下角的“提交”。



1. 查看购买的弹性公网IP

返回到控制台后，点击刷新按钮可以看到已购买的2个弹性IP。 

### 任务二：开通MapReduce服务

1. 打开MRS服务

在控制台展开“服务列表”，在列表中点击“EI企业智能”分类里面的“MapReduce服务”。



申请集群

在MRS控制台页面点击右上角的“购买集群”。



集群配置：软件配置

选择“**自定义购买**”进入软件配置界面。区域选“华北-北京四”。



集群名称“mrs\_rta”（可以自定义），集群版本选则“MRS 1.9.2”，集群类型选择“流式集群”并勾选所有组件。

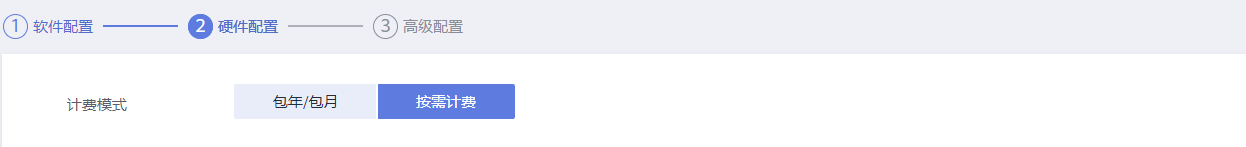


点击右下角的“下一步”。



集群配置：硬件配置

进入硬件配置界面，计费模式选择“按需计费”。



可用区默认，选择虚拟私有云（如果还没有虚拟私有云，点击“查看虚拟私有云”进入虚拟私有云控制台，参照**附录5.1**进行创建，创建后回到MRS控制台并点击虚拟私有云后面的刷新按钮），子网会自动选定，安全组“自动创建”，弹性公网IP处点击下拉选择一个前面购买的弹性公网IP（此处IP是MRS集群管理界面的IP）。



CPU架构选择“鲲鹏计算”，**关闭Master节点实例数量中的“集群高可用”，**流式Core节点的实例数量改为“1”。



点击右下角的“下一步”。



集群配置：高级配置

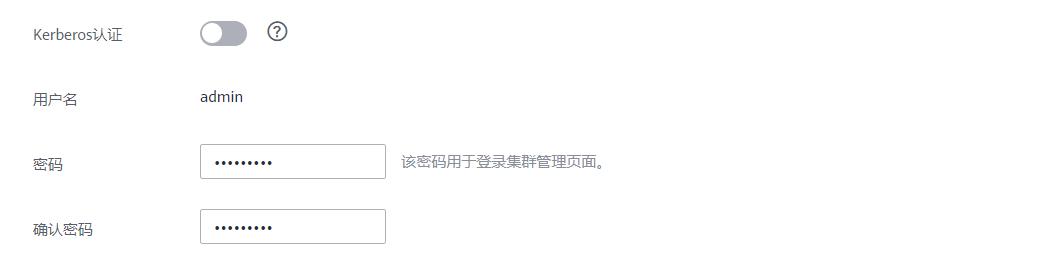
标签、弹性伸缩、引导操作保持默认。



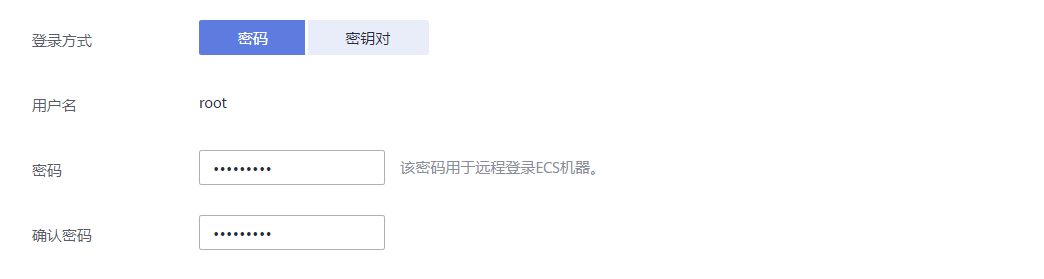
告警选择“关闭”，委托、数据盘加密默认即可。



**关闭“Kerberos认证”**，按规则设置密码（该密码用于登录集群管理页面）。



登录方式选择“密码”，输入密码（此处为登录MRS集群节点ECS机器的面）。



通信安全授权，勾选“确认授权”复选框。



点击右下角的“立即购买”。



查看购买的集群

点击“返回集群列表”。



可以看到集群状态为“启动中”（集群创建需要10分钟左右的时间）。



状态变为“运行中”时，集群创建成功。



给集群的master节点绑定弹性IP

在现有集群列表中点击集群名称，进入集群页面。



在集群页面点击“节点管理”，点击下箭头展开master\_node\_default\_group节点组，点击master的节点名称进入master节点（华为云只在master节点上安装了组件的客户端，登录其他节点无法操作组件）。



选择“弹性公网IP”，点击“绑定弹性公网IP”（弹性IP在前序步骤中已经购买）。



选择一个弹性公网IP后点击“确定”。



点击右上角的刷新按钮后可以看到已经绑定的IP。



弹性公网IP绑定完毕，后面可以使用PuTTY等远程工具通过此IP登录集群的master节点服务器进行相关的操作。

修改安全组规则

在master节点界面选择“安全组”，点击下箭头展开安全组，点击“更改安全组规则”（配置好IP后需要修改安全组规则，否则无法登陆到master节点服务器）。



进入安全组后选择“入方向规则”，点击“添加规则”。



优先级为“1”（优先级范围1~100，数字越小级别越高），策略“允许”，协议端口处选择“TCP”>“基本协议”>“全部放通”，然后点击确定。



MRS服务到此购买并配置完成。

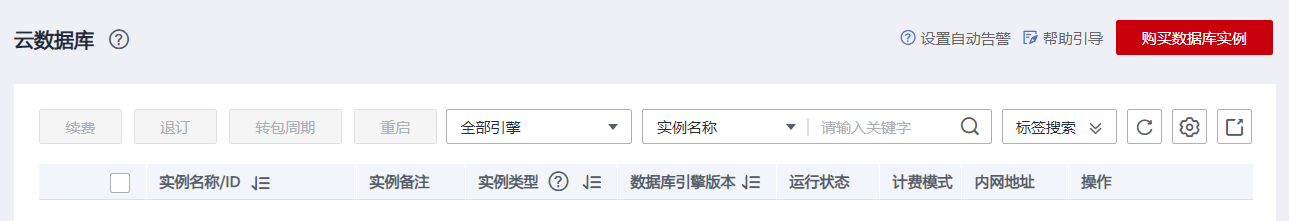
### 任务三：开通云数据库服务RDS

1. 进入RDS服务

在控制台展开“服务列表”，在列表中点击“数据库”分类里面的“云数据库RDS”。



在实例管理中，点击右上角的“购买数据库实例”。



设置RDS配置信息

计费模式选择“按需计费”，区域选择“华北-北京四”。



实例名称“rds-rta”（可以自定义），数据库引擎“MySQL”，数据库版本“5.7”，实例类型“**单机**”，存储类型、可用区、时区默认。



性能规格选择“鲲鹏通用增强型”>“2 vCPUs | 4 GB”，存储空间、磁盘加密默认。

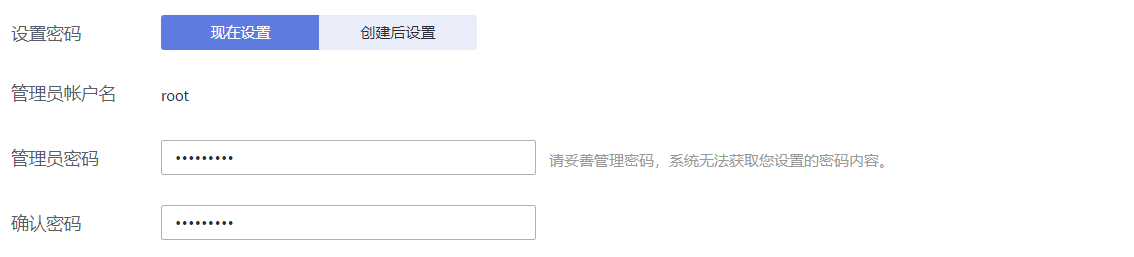


选择虚拟私有云和子网（使用与MRS相同的虚拟私有云和子网），安全组、端口默认即可。

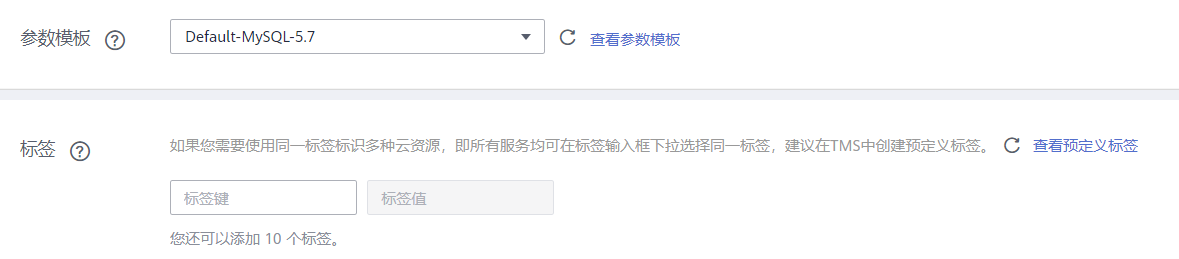


注意：此处选择的安全组与MRS安全组一致。

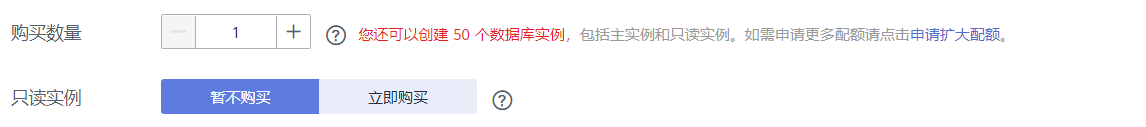
设置密码。



参数模板和标签默认。



购买数量“1”，只读实例“暂不购买”。



点击右下角的“立即购买”。



查看详情没有问题后点击右下角的“提交”。



查看购买的RDS实例

点击“返回云数据库RDS列表”。



可以看到实例状态为“创建中”（创建过程需要几分钟的时间）。



等待实例的状态变为“运行中”则实例创建完成。



配置MySQL实例

在实例管理界面，点击刚创建的实例名称（如果浏览器关闭了，可以通过控制台-服务列表-数据库-中再找到云数据库RDS进入）。

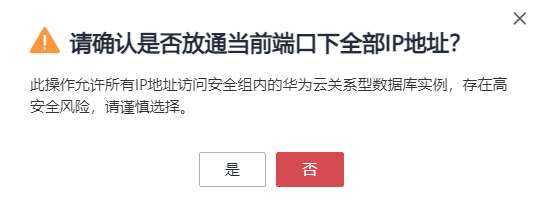


进入RDS实例页面，点击左侧的“连接管理”，选择“内网连接”可以查看到MySQL的内网IP地址，此IP地址会在后面步骤中用到（点后面的方形按钮可以复制）。



在内网连接界面向下滚动页面，在“安全组规则”中，选择“入方向规则”，然后点击“一键添加”。 

弹出对话框点击“是”。



配置完成。

### 任务四：开通数据湖探索服务（DLI）

1. 进入数据库探索DLI服务

在控制台展开“服务列表”，在列表中点击“EI企业智能”分类里面的“数据湖探索DLI”。



如果出现授权，同意授权即可。



队列购买

队列是DLI的计算资源，队列类型有两种：[SQL队列和通用队列](https://console.huaweicloud.com/dli/?agencyId=080b01efd58025211f57c00db7833d24&region=cn-east-2&locale=zh-cn#/resource/queues)；SQL队列支持提交Spark SQL作业，通用队列支持Spark程序、Flink SQL、Flink Jar作业。

点击右上角的“购买队列”按钮



计费模式选择“按需计费”，区域选择“华北-北京四”。



队列名称“queue\_flink”（可以自定义），队列类型选择“通用队列”，**勾选专属资源模式**，AZ策略选择“单AZ”，CPU架构“鲲鹏”，队列规格“16CUs”，高级配置、标签默认。



点击右下角的“立即购买”。



确认配置后，点击右下角的“提交”。



队列创建成功。



服务授权

在“全局配置”中选择“服务授权”，勾选“VPC Administrator”（创建跨源连接需要）。



点击“更新委托权限”按钮。



配置跨源链接

DLI增强型跨源连接底层采用对等连接，直接打通DLI集群与数据源的vpc网络，通过点对点的方式实现数据互通。后续实验中DLI要读取MRS的Kafka中的数据，需要将DLI服务的网络和MRS的网络打通，所以需要进行跨源连接的配置。

在DLI控制台页面，点击“跨源连接”菜单，在“增强型跨源”页签下点击 “创建”按钮。



输入连接名称“DLI\_RDS”(可自定义)，绑定队列选择上一步创建的DLI队列，选择虚拟私有云和子网（**均与MRS服务相同**），主机信息、标签默认，点击“确定”。



跨源连接配创建完成。



注意：因为我们购买的RDS-MySQL服务和MRS服务是用的同一个VPC网络，所以这里创建一个连接就能够实现DLI服务读取MRS中Kafka的数据，同时DLI也能把数据发送到MySQL中。如果MySQL和MRS不在同一个VPC网络，那么这里需要创建两个连接。

查看对等连接

增强型跨源连接会自动创建对等连接来打通网络，所以必须保证对等连接创建成功后面才能发挥其作用。

点击跨源连接的连接名称，进入到连接的详情页面，可以看到连接状态为“CREATING”，表示对等连接正在创建，创建过程需要几分钟的时间。



点击右上角的刷新按钮，等到连接状态变为“ACTIVE”则对等连接创建成功同时获取到了对等连接ID。



注：若未能成功创建对等连接，则需要解绑队列后重新绑定队列。

### 任务五：开通数据迁移服务（CDM）

在实时流数据可视化的部分中，DLV服务要读取MySQL中的数据，需要使用云数据迁移服务CDM。

1. 进入CDM服务

在控制台展开“服务列表”，在列表中点击“迁移”分类里面的“云数据迁移CDM”。 

在集群管理界面，点击右上角的“购买云数据迁移服务”。

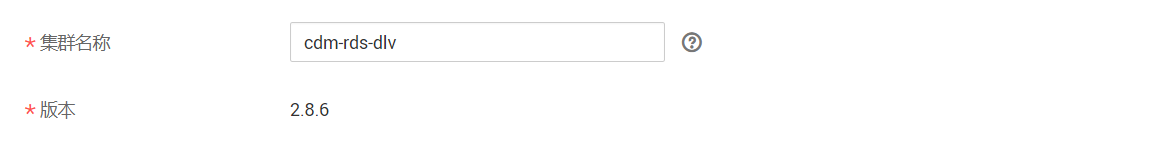


服务选型及购买

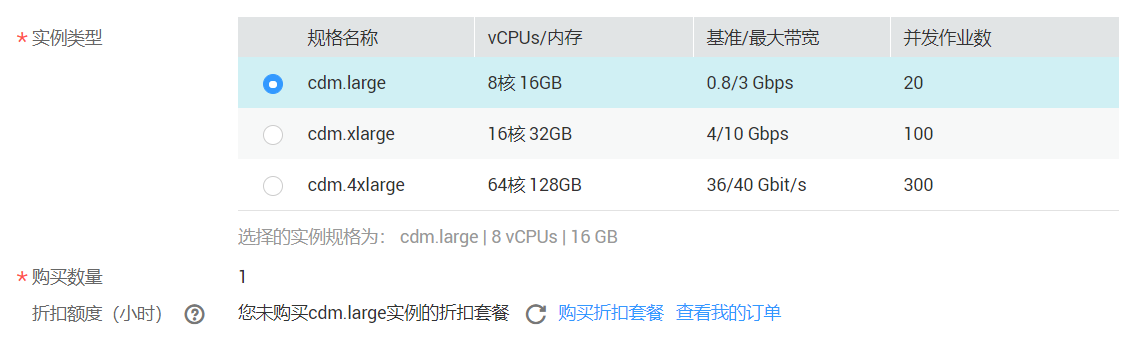
当前区域选择“华北-北京四”（保持和MRS服务在同一区域），可用区默认。



集群名称“cdm-rds-dlv”（可以自定义），版本默认。



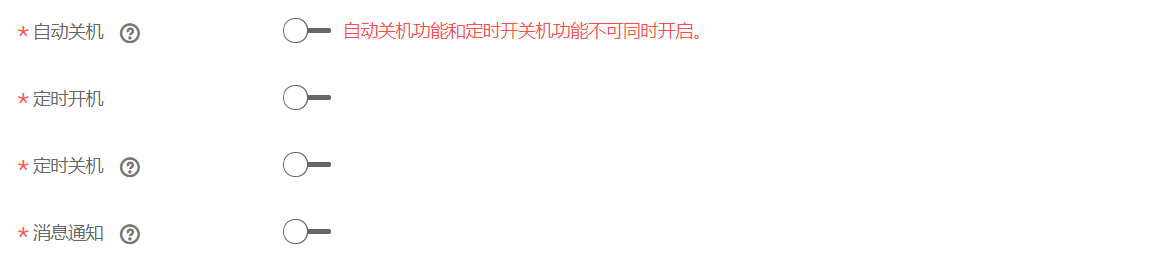
实例类型选择“cdm.large”，购买数量默认“1”



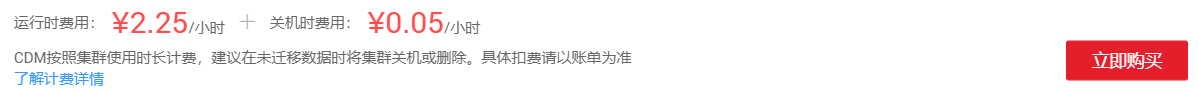
虚拟私有云、子网、安全组与MRS的配置相同。



其他配置默认。



点击右下角的“立即购买”。



确认规格，若有问题点击“上一步”修改，确认无误后点击“提交”。



查看集群

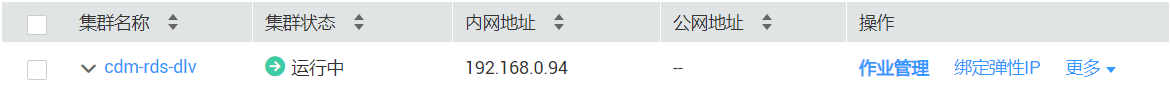
点击“返回集群管理页面”



集群创建需要几分钟的时间，请耐心等待。



集群状态变为“运行中”时，集群创建成功。



四：实验内容及步骤、实验的详细记录、实验结果分析

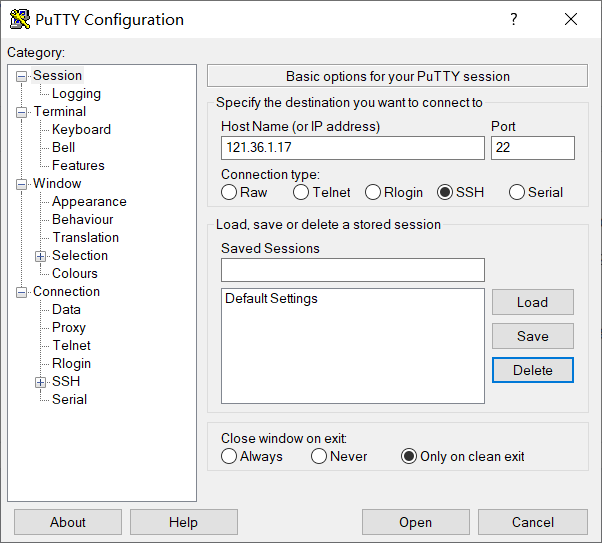
**请附上实验过程截图（截图需包含指令）以及必要的文字分析**

## 实验任务

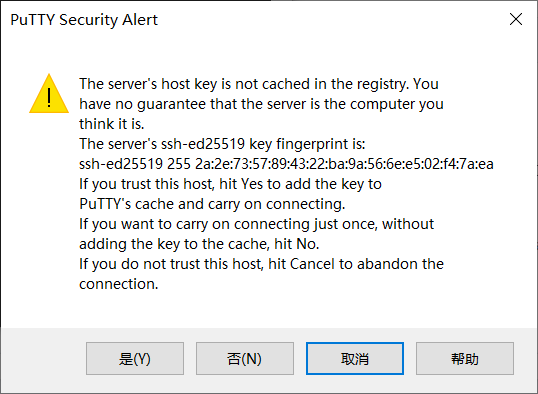
### 任务一：Python脚本生成测试数据(20’)

1. 登录MRS的master节点服务器

打开Putty，输入前面为MRS的master节点绑定的公网IP，点击Open



弹出的对话框点击“是”。



输入用户名和密码登录到MRS服务的master节点（注：密码不显示任何字符，输入后回车即可）。



编写Python脚本

进入/opt/client/目录，使用vi命令编写Python脚本：autodatagen.py

cd /opt/client/

vim autodatagen.py

复制下面的脚本代码，粘贴到PuTTY的vim中（在PuTTY中右键即为粘贴）：

#coding:utf-8

###########################################

# rowkey：随机的两位数 + 当前时间戳，并要确保该rowkey在表数据中唯一。

# 列定义：行健，用户名，年龄，性别，商品ID，价格，门店ID，购物行为，电话，邮箱，购买日期

#421564974572,Sgxrp,20,woman,152121,297.64,313015,scan,15516056688,JbwLDQmzwP@qq.com,2019-08-01

#601564974572,Lbeuo,43,man,220902,533.13,313016,pv,15368953106,ezfrJSluoR@163.com,2019-08-05###########################################

import random

import string

import sys

import time

# 大小写字母

alphabet\_upper\_list = string.ascii\_uppercase

alphabet\_lower\_list = string.ascii\_lowercase

# 随机生成指定位数的字符串

def get\_random(instr, length):

# 从指定序列中随机获取指定长度的片段并组成数组，例如:['a', 't', 'f', 'v', 'y']

res = random.sample(instr, length)

# 将数组内的元素组成字符串

result = ''.join(res)

return result

# 放置生成的并且不存在的rowkey

rowkey\_tmp\_list = []

# 制作rowkey

def get\_random\_rowkey():

import time

pre\_rowkey = ""

while True:

# 获取00~99的两位数字，包含00与99

num = random.randint(00, 99)

# 获取当前10位的时间戳

timestamp = int(time.time())

# str(num).zfill(2)为字符串不满足2位，自动将该字符串补0

pre\_rowkey = str(num).zfill(2) + str(timestamp)

if pre\_rowkey not in rowkey\_tmp\_list:

rowkey\_tmp\_list.append(pre\_rowkey)

break

return pre\_rowkey

# 创建用户名

def get\_random\_name(length):

name = string.capwords(get\_random(alphabet\_lower\_list, length))

return name

# 获取年龄

def get\_random\_age():

return str(random.randint(18, 60))

# 获取性别

def get\_random\_sex():

return random.choice(["woman", "man"])

# 获取商品ID

def get\_random\_goods\_no():

goods\_no\_list = ["220902", "430031", "550012", "650012", "532120","230121","250983", "480071", "580016", "950013", "152121","230121"]

return random.choice(goods\_no\_list)

# 获取商品价格（浮点型）

def get\_random\_goods\_price():

# 随机生成商品价格的整数位，1~999的三位数字，包含1与999

price\_int = random.randint(1, 999)

# 随机生成商品价格的小数位，1~99的两位数字，包含1与99

price\_decimal = random.randint(1, 99)

goods\_price = str(price\_int) +"." + str(price\_decimal)

return goods\_price

# 获取门店ID

def get\_random\_store\_id():

store\_id\_list = ["313012", "313013", "313014", "313015", "313016","313017","313018", "313019", "313020", "313021", "313022","313023"]

return random.choice(store\_id\_list)

# 获取购物行为类型

def get\_random\_goods\_type():

goods\_type\_list = ["pv", "buy", "cart", "fav","scan"]#点击、购买、加购、收藏、浏览

return random.choice(goods\_type\_list)

# 获取电话号码

def get\_random\_tel():

pre\_list = ["130", "131", "132", "133", "134", "135", "136", "137", "138", "139", "147", "150",

"151", "152", "153", "155", "156", "157", "158", "159", "186", "187", "188"]

return random.choice(pre\_list) + ''.join(random.sample('0123456789', 8))

# 获取邮箱名

def get\_random\_email(length):

alphabet\_list = alphabet\_lower\_list + alphabet\_upper\_list

email\_list = ["163.com", "126.com", "qq.com", "gmail.com","huawei.com"]

return get\_random(alphabet\_list, length) + "@" + random.choice(email\_list)

# 获取商品购买日期（统计最近7天数据）

def get\_random\_buy\_time():

buy\_time\_list = ["2019-08-01", "2019-08-02", "2019-08-03", "2019-08-04", "2019-08-05", "2019-08-06", "2019-08-07"]

return random.choice(buy\_time\_list)

# 生成一条数据

def get\_random\_record():

return get\_random\_rowkey() + "," + get\_random\_name(

5) + "," + get\_random\_age() + "," + get\_random\_sex() + "," + get\_random\_goods\_no() + ","+get\_random\_goods\_price()+ "," +get\_random\_store\_id()+ "," +get\_random\_goods\_type() + ","+ get\_random\_tel() + "," + get\_random\_email(

10) + "," +get\_random\_buy\_time()

# 获取随机整数用于休眠

def get\_random\_sleep\_time():

return random.randint(5, 10)

# 将记录写到文本中

def write\_record\_to\_file():

# 覆盖文件内容，重新写入

f = open(sys.argv[1], 'w')

i = 0

while i < int(sys.argv[2]):

record = get\_random\_record()

f.write(record)

# 换行写入

f.write('\n')

i += 1

f.close()

if \_\_name\_\_ == "\_\_main\_\_":

write\_record\_to\_file()

输入 :wq 保存并退出vim编辑器。

创建存放测试数据的目录

使用mkdir命令在/tmp下创建目录flume\_spooldir，我们把Python脚本模拟生成的数据放到此目录下，后面Flume就监控这个文件下的目录，以读取数据。

mkdir /tmp/flume\_spooldir

执行脚本测试

执行Python命令，测试生成100条数据

python autodatagen.py "/tmp/flume\_spooldir/test.txt" 100

使用more命令查看生成的数据（enter键向下一行，space键向下一屏，b向上一屏，q键退出）。

more /tmp/flume\_spooldir/test.txt

### 任务二：配置Kafka(10’)

1. 设置环境变量

使用Putty登录MRS的master节点服务器后，首先使用source命令进行环境变量的设置使得相关命令可用。

source /opt/client/bigdata\_env

在kafka中创建topic

执行如下命令创建topic（替换实际Zookeeper的IP，如何获取IP请参见附录步骤）。

kafka-topics.sh --create --zookeeper **192.168.0.183**:2181/kafka --partitions 1 --replication-factor 1 --topic fludesc

注意命令执行时所在目录和命令大小写（可以使用绝对路径：/opt/client/Kafka/kafka/bin/kafka-topics.sh）

查看topic信息

kafka-topics.sh --list --zookeeper 192.168.0.183:2181/kafka



### 任务三：安装Flume客户端(10’)

1. 进入MRS Manager集群管理

在 mrs\_rta集群页面，点击“前往Manager”。



输入用户名和密码，点击“登录”。



进入到MRS Manager界面。



1. 下载Flume客户端

在MRS Manager集群管理界面，点击“服务管理”，点击“Flume”，进入Flume服务。



点击“下载客户端”按钮。



点击“确定”后等待下载。



下载完成后会有弹出框提示下载到哪一台服务器上（这台机器就是master节点），路径就是/tmp/MRS-client。

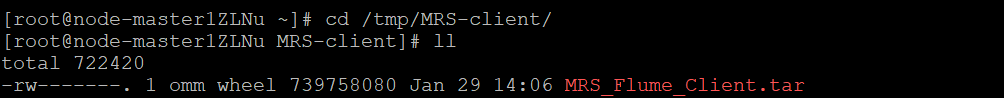


校验下载的客户端文件包

使用PuTTY登录到master节点服务器上，进入/tmp/MRS-client目录。

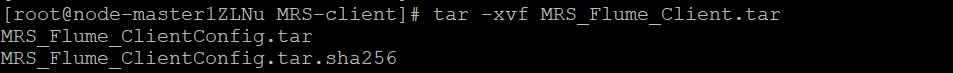
cd /tmp/MRS-client/

ll



执行以下命令，解压压缩包获取校验文件与客户端配置包

tar -xvf MRS\_Flume\_Client.tar



执行命令，校验文件包。

sha256sum -c MRS\_Flume\_ClientConfig.tar.sha256



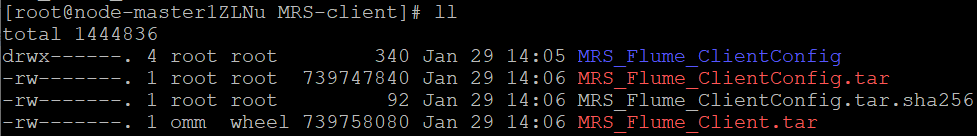
界面显示如上信息，表明文件包校验成功。

安装Flume运行环境

解压“MRS\_Flume\_ClientConfig.tar”文件。

tar -xvf MRS\_Flume\_ClientConfig.tar

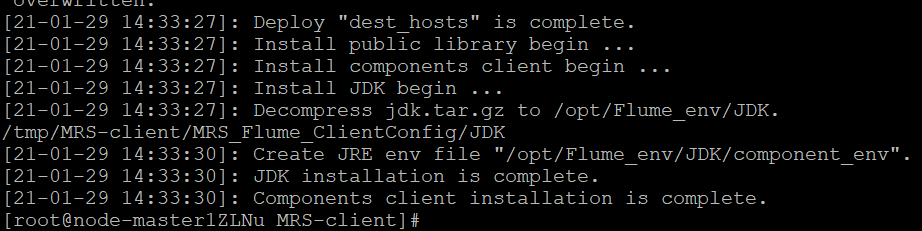
查看解压后文件。



安装客户端运行环境到目录“/opt/Flume\_env”（安装时自动生成目录）。

sh /tmp/MRS-client/MRS\_Flume\_ClientConfig/install.sh /opt/Flume\_env

如有以下输出表示客户端运行环境安装成功。



执行命令配置环境变量。

source /opt/Flume\_env/bigdata\_env

安装Flume客户端

执行命令解压Flume客户端文件。

cd /tmp/MRS-client/MRS\_Flume\_ClientConfig/Flume

tar -xvf FusionInsight-Flume-1.6.0.tar.gz

安装Flume到目录“/opt/FlumeClient”（安装时自动生成目录，-d：表示Flume客户端安装路径）。

sh /tmp/MRS-client/MRS\_Flume\_ClientConfig/Flume/install.sh -d /opt/FlumeClient

系统显示以上结果表示客户端运行环境安装成功。

重启Flume服务

执行以下命令重启Flume的服务。

cd /opt/FlumeClient/fusioninsight-flume-1.6.0

sh bin/flume-manage.sh restart

服务重启成功，安装结束！

### 任务四：配置Flume采集数据(20’)

1. 修改配置文件

进入Flume安装目录

cd /opt/FlumeClient/fusioninsight-flume-1.6.0/

在conf目录下编辑文件properties.properties

vi conf/properties.properties

加入如下内容（注意修改后面说明的内容）:

client.sources = s1

client.channels = c1

client.sinks = sh1

# the source configuration of s1

client.sources.s1.type = spooldir

client.sources.s1.spoolDir = /tmp/flume\_spooldir

client.sources.s1.fileSuffix = .COMPLETED

client.sources.s1.deletePolicy = never

client.sources.s1.trackerDir = .flumespool

client.sources.s1.ignorePattern = ^$

client.sources.s1.batchSize = 1000

client.sources.s1.inputCharset = UTF-8

client.sources.s1.deserializer = LINE

client.sources.s1.selector.type = replicating

client.sources.s1.fileHeaderKey = file

client.sources.s1.fileHeader = false

client.sources.s1.basenameHeader = true

client.sources.s1.basenameHeaderKey = basename

client.sources.s1.deserializer.maxBatchLine = 1

client.sources.s1.deserializer.maxLineLength = 2048

client.sources.s1.channels = c1

# the channel configuration of c1

client.channels.c1.type = memory

client.channels.c1.capacity = 10000

client.channels.c1.transactionCapacity = 1000

client.channels.c1.channlefullcount = 10

client.channels.c1.keep-alive = 3

client.channels.c1.byteCapacityBufferPercentage = 20

# the sink configuration of sh1

client.sinks.sh1.type = org.apache.flume.sink.kafka.KafkaSink

client.sinks.sh1.kafka.topic = **fludesc**

client.sinks.sh1.flumeBatchSize = 1000

client.sinks.sh1.kafka.producer.type = sync

client.sinks.sh1.kafka.bootstrap.servers = **192.168.0.226**:9092

client.sinks.sh1.kafka.security.protocol = PLAINTEXT

client.sinks.sh1.requiredAcks = 0

client.sinks.sh1.channel = c1

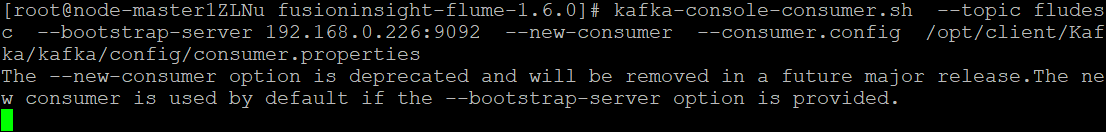
**client.sinks.sh1.kafka.topic** 的值为kafka中创建的topic的值；**client.sinks.sh1.kafka.bootstrap.servers**的值为Kafka实例Broker的IP和端口（如何查看Kafka的IP，请参见附录步骤）。

输入 :wq 保存后退出编辑器。

创建消费者消费kafka中的数据

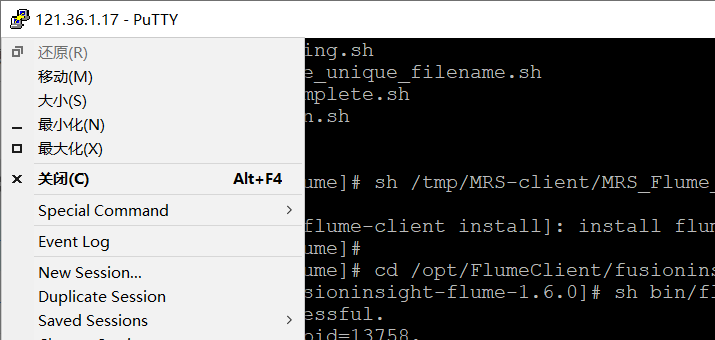
使用PuTTY登录master节点后，执行命令（此处bootstrap-server的ip对应的是Kafka的Broker的IP）：

kafka-console-consumer.sh --topic fludesc --bootstrap-server **192.168.0.226**:9092 --new-consumer --consumer.config /opt/client/Kafka/kafka/config/consumer.properties

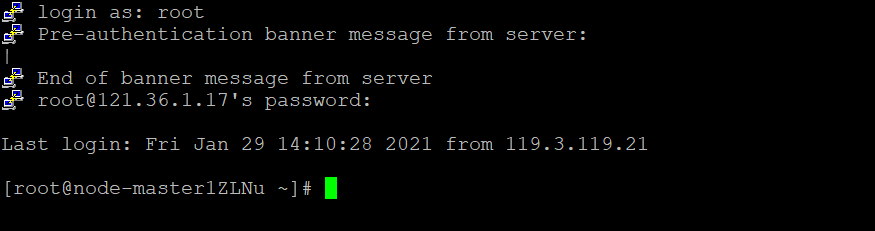


执行以上命令后，需要**新开一个PuTTY会话窗口**。

点击PuTTY标题栏左侧的图标，在弹出的菜单中选择“Duplicate Session”复制会话。



在新弹出的会话窗口中输入用户名和密码登录。

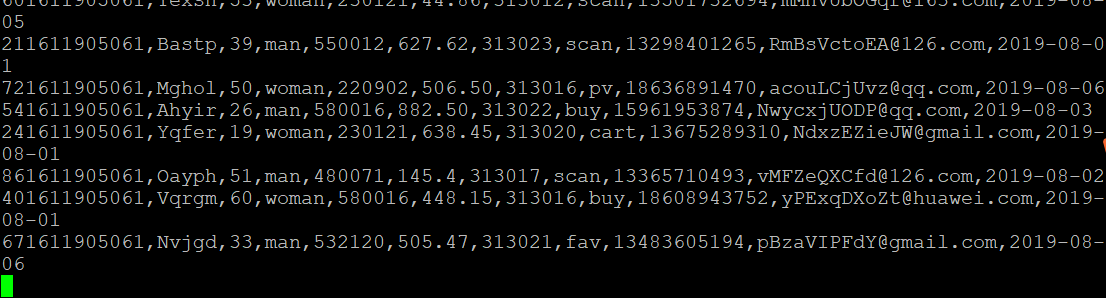


进入Python脚本所在目录，执行python脚本，再生成一份数据。

cd /opt/client/

python autodatagen.py "/tmp/flume\_spooldir/test.txt" 100

查看原窗口，可以看到已经消费出了数据：



有数据产生，表明Flume到Kafka目前是打通的。

测试完毕，在新打开的窗口输入exit关闭窗口，在原窗口输入 Ctrl+c退出进程。

### 任务五：MySQL中准备结果表与维度表数据(10’)

1. 登录MySQL的DAS管理平台。

在控制台进入到云数据库RDS实例管理界面，点击实例后面的“登录”按钮。



输入用户名root和密码，勾选“记住密码”，开启“定时采集”和“SQL执行记录”，然后点击“测试连接”，成功后点击“登录”按钮。



创建数据库

点击“新建数据库”。



输入名称“rds\_desc”，字符集选择utf8，点击“确定”。



进入SQL界面

点击数据库后面的“SQL查询”。



进入到SQL执行界面。



创建维度表并插入数据

清除查询器中原有内容，复制下面的SQL语句粘贴到SQL查询中：

-- ----------------------------

-- 创建商品信息纬度表

-- ----------------------------

DROP TABLE IF EXISTS `desc\_goods\_info`;

CREATE TABLE `desc\_goods\_info` (

`goods\_no` varchar(30) NOT NULL,

`goods\_name` varchar(30) DEFAULT NULL

) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------

-- 插入商品信息样例数据

-- ----------------------------

INSERT INTO `desc\_goods\_info` VALUES ('220902', '杭州丝绸');

INSERT INTO `desc\_goods\_info` VALUES ('430031', '西湖龙井');

INSERT INTO `desc\_goods\_info` VALUES ('550012', '西湖莼菜');

INSERT INTO `desc\_goods\_info` VALUES ('650012', '张小泉剪刀');

INSERT INTO `desc\_goods\_info` VALUES ('532120', '塘栖枇杷');

INSERT INTO `desc\_goods\_info` VALUES ('230121', '临安山核桃');

INSERT INTO `desc\_goods\_info` VALUES ('250983', '西湖藕粉');

INSERT INTO `desc\_goods\_info` VALUES ('480071', '千岛湖鱼干');

INSERT INTO `desc\_goods\_info` VALUES ('580016', '天尊贡芽');

INSERT INTO `desc\_goods\_info` VALUES ('950013', '叫花童鸡');

INSERT INTO `desc\_goods\_info` VALUES ('152121', '火腿蚕豆');

INSERT INTO `desc\_goods\_info` VALUES ('230121', '杭州百鸟朝凤');

-- ----------------------------

-- 创建门店信息纬度表

-- ----------------------------

DROP TABLE IF EXISTS `desc\_store\_info`;

CREATE TABLE `desc\_store\_info` (

`store\_id` varchar(50) NOT NULL,

`store\_name` varchar(50) DEFAULT NULL

) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------

-- 插入门店信息样例数据

-- ----------------------------

INSERT INTO `desc\_store\_info` VALUES ('313012', '莫干山店');

INSERT INTO `desc\_store\_info` VALUES ('313013', '定安路店');

INSERT INTO `desc\_store\_info` VALUES ('313014', '西湖银泰店');

INSERT INTO `desc\_store\_info` VALUES ('313015', '天目山店');

INSERT INTO `desc\_store\_info` VALUES ('313016', '凤起路店');

INSERT INTO `desc\_store\_info` VALUES ('313017', '南山路店');

INSERT INTO `desc\_store\_info` VALUES ('313018', '西溪湿地店');

INSERT INTO `desc\_store\_info` VALUES ('313019', '传媒学院店');

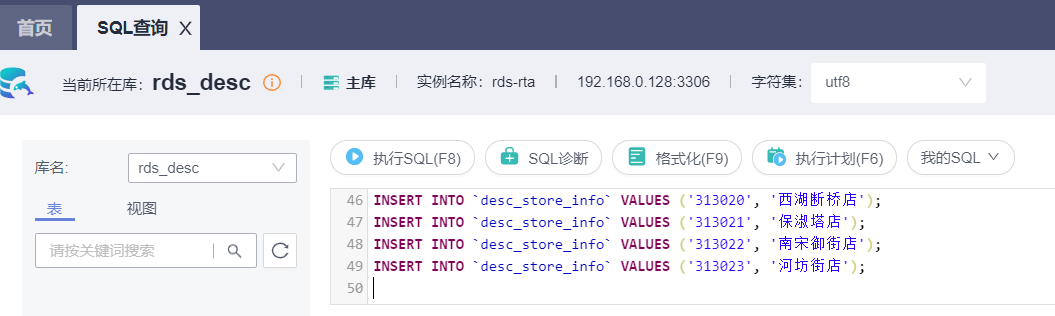
INSERT INTO `desc\_store\_info` VALUES ('313020', '西湖断桥店');

INSERT INTO `desc\_store\_info` VALUES ('313021', '保淑塔店');

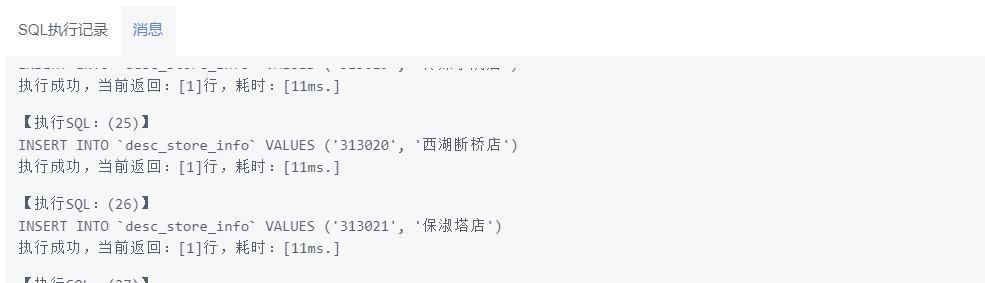
INSERT INTO `desc\_store\_info` VALUES ('313022', '南宋御街店');

INSERT INTO `desc\_store\_info` VALUES ('313023', '河坊街店');

点击“执行SQL”执行上面的语句：



执行成功后可以在下面看到执行消息：



创建Flink作业的结果表

Flink执行的结果数据是存储在MySQL中，现在需要提前把表创建好。

步骤同上，清除SQL窗口中原有代码（**若不清除原有代码则需要选中本次要执行的代码后再执行**），粘贴并执行下面建表语句：

-- ----------------------------

-- 创建商品总销售额表--goods\_amount\_count

-- ----------------------------

DROP TABLE IF EXISTS `goods\_amount\_count`;

CREATE TABLE `goods\_amount\_count` (

`amount\_total` float NOT NULL,

`sale\_date` date PRIMARY KEY

) ENGINE=InnoDB DEFAULT CHARSET=utf8;

-- ----------------------------

-- 创建销售总额前5的门店排行表--amount\_store\_rank

-- ----------------------------

DROP TABLE IF EXISTS `amount\_store\_rank`;

CREATE TABLE `amount\_store\_rank` (

`store\_id` int PRIMARY KEY,

`store\_name` varchar(50) DEFAULT NULL,

`amount\_total` float DEFAULT NULL

) ENGINE=InnoDB DEFAULT CHARSET=utf8;

执行成功后，点击左侧的刷新按钮可以看到已经创建的表：



### 任务六：使用DLI中的Flink作业进行数据分析(20’)

1. 创建Flink作业

进入数据湖探索服务（DLI）的控制台，点击左侧“作业管理”菜单，选择“Flink作业”，然后点击右上角的“创建作业”。 

类型选择“Flink SQL”，名称输入“desc\_order\_count”（可以自定义），模板、标签默认，点击“确定”。



编辑Flink作业的SQL脚本

复制下面的脚本到编辑框中（需要按照后面的提示修改相关参数）。

/\*\*

\* 该示例为CS服务使用对等连接对接用户在华为云自有的Kafka集群的SQL模板。数据的输入源为用户的Kafka集群提供

\* >>>>>>>>>请务必确保您的Kafka集群中已创建了您配置的Kafka主题<<<<<<<<<<

\*

\* >>>>>样例输入<<<<<

\* 流名: \*kafka\_source(id,user\_name,age,gender,goods\_no,goods\_price,store\_id,shopping\_type,tel,email,shopping\_date):

\*行健，用户名，年龄，性别，商品ID，价格，门店ID，购物行为，电话，邮箱，购买日期

\*591564404906,Gkisf,59,man,550012,23.45,423012,fav,15274039826,UsPZNdxnJX@163.c\*om,2019-07-01

\*761564404906,Wupsa,55,man,250983,25.05,423013,scan,13730586914,AKVMdPveQX@gmai\*l.com,2019-07-02

\*\*/

/\*\***kafka输入流**\*\*/

CREATE SOURCE STREAM kafka\_source (id STRING,use\_rname STRING, age int,gender STRING,goods\_no STRING,goods\_price Float,store\_id int,shopping\_type STRING,tel STRING,email STRING,shopping\_date Date)

WITH (

type = "kafka",

kafka\_bootstrap\_servers = "**192.168.0.226**:9092", /\*\* kafka的连接端口，需要确保能连通（可以使用对等连接的方式打通CS集群和kafka集群）\*\*/

kafka\_group\_id = "sourcegroup1", /\*\* group id\*\*/

kafka\_topic = "**fludesc**", /\*\*读取的topic\*\*/

encode = "csv", /\*\* 解码格式，支持json/csv\*\*/

field\_delimiter = ","

);

/\*

\* >>>>>1-商品总的销售额（按天统计）数据输出<<<<<

\* **流名: goods\_amount\_count**

\*

\*\*/

CREATE SINK STREAM goods\_amount\_count(

amount\_total Float,/\* 销售额\*/

sale\_date date /\* 销售日期\*/

)

WITH (

type = "rds",

username = "root", /\*MySQL的用户名默认是root \*/

password = "**填写密码**", /\*创建MySQL实例设置的密码\*/

db\_url = "mysql://**192.168.0.181**:3306/rds\_desc", /\*此处IP改成MySQL的内网IP，端口号不需要改，默认就是3306，IP可以在MySQL页面查看\*/

table\_name = "goods\_amount\_count",

primary\_key = "sale\_date"

);

/\*

\* >>>>>2-销售总额前5的门店排行数据输出<<<<<

\* **流名: amount\_store\_rank**

\*

\*\*/

CREATE SINK STREAM amount\_store\_rank(

store\_id int,/\* 门店ID\*/

store\_name STRING,/\* 门店名称\*/

amount\_total Float /\* 销售额\*/

)

WITH (

type = "rds",

username = "root", /\*MySQL的用户名默认是root \*/

password = "**填写密码**", /\*创建MySQL实例设置的密码\*/

db\_url = "mysql://**192.168.0.181**:3306/rds\_desc", /\*此处IP改成MySQL的内网IP，端口号不需要改，默认就是3306，IP可以在MySQL页面查看\*/

table\_name = "amount\_store\_rank",

primary\_key = "store\_id"

);

/\*

创建数据维表1，用于和输入流连接，实现字段回填

本实验中用到的是关联维度表门店信息表，根据流中的门店id,关联出门店的名称

\*\*/

CREATE TABLE table\_store\_info (

store\_id STRING,

store\_name STRING

)

WITH (

type = "rds",

region = "**cn-north-4**",/\* 服务所在的区域，例如华北-北京四为：cn-north-4\*/

username = "root", /\*MySQL的用户名默认是root \*/

password = "**填写密码**", /\*创建MySQL实例设置的密码\*/

db\_url = "mysql://**192.168.0.181**:3306/rds\_desc", /\*此处IP改成MySQL的内网IP，端口号不需要改，默认就是3306，IP可以在MySQL页面查看\*/

table\_name = "desc\_store\_info"/\* 关联的维度表名称\*/

);

/\*

创建数据维表2，用于和输入流连接，实现字段回填

本实验中用到的是关联维度表商品信息表，根据流中的商品id,关联出商品的名称

\*\*/

CREATE TABLE table\_goods\_info (

goods\_no STRING,

goods\_name STRING

)

WITH (

type = "rds",

region = "**cn-north-4**",/\* 服务所在的区域，例如华北-北京四为：cn-north-4\*/

username = "root", /\*MySQL的用户名默认是root \*/

password = "**填写密码**", /\*创建MySQL实例设置的密码\*/

db\_url = "mysql://**192.168.0.181**:3306/rds\_desc", /\*此处IP改成MySQL的内网IP，端口号不需要改，默认就是3306，IP可以在MySQL页面查看\*/

table\_name = "desc\_goods\_info"/\* 关联的维度表名称\*/

);

/\* 1-商品总的销售额数据插入 \*/

INSERT INTO goods\_amount\_count

SELECT sum(goods\_price) as amount\_total,shopping\_date as sale\_date

FROM kafka\_source

WHERE shopping\_type = 'buy'

group BY shopping\_date;

/\* 2-销售总额前5的门店排行数据插入 \*/

INSERT INTO amount\_store\_rank

SELECT t1.store\_id,t2.store\_name,sum(t1.goods\_price) as amount\_total

FROM kafka\_source t1 left join table\_store\_info as t2

on t1.store\_id=t2.store\_id

WHERE t1.shopping\_type = 'buy'

group BY t1.store\_id,t2.store\_name;

上面SQL代码中的需要修改的参数说明如下。

kafka输入流中：

**kafka\_bootstrap\_servers**的值修改为MRS集群中Kafka的Broker的IP和端口号；

**kafka\_topic**的值修改为kafka中创建的主题名称；

goods\_amount\_count流中：

**password**修改为MySQL的登录密码；

**db\_url**修改为MySQL的**内网IP**；

**amount\_store\_rank流中：**

**password**修改为MySQL的登录密码；

**db\_url**修改为MySQL的**内网IP**；

**table\_store\_info和table\_goods\_info维度表**：

**region**的值为MySQL服务所在的区，如何查找对应的值，参见附录步骤；

**password**修改为MySQL的登录密码；

**db\_url**修改为MySQL的**内网IP**；



编辑完后点击“语义校验”，如果校验无错误则进行下一步骤操作，如果有误则根据报错信息对SQL脚本进行修改后重新校验。

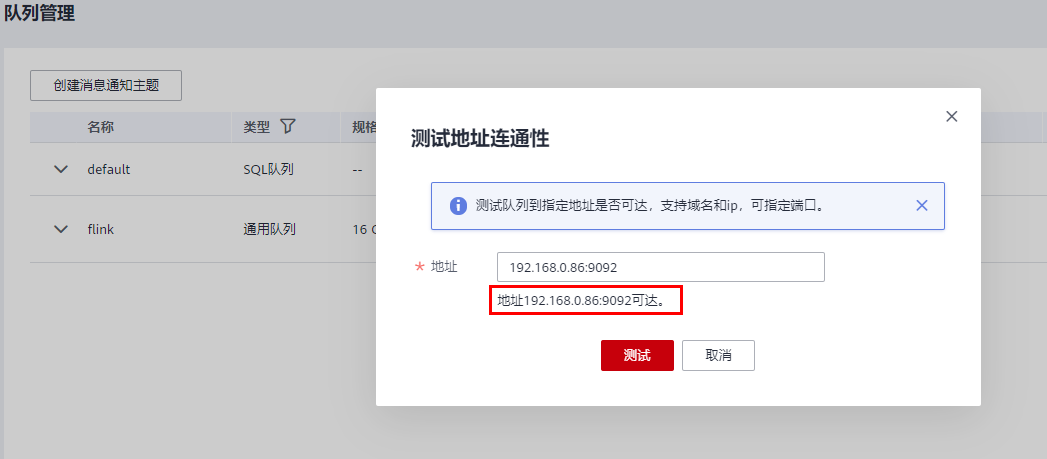


测试网络连通性

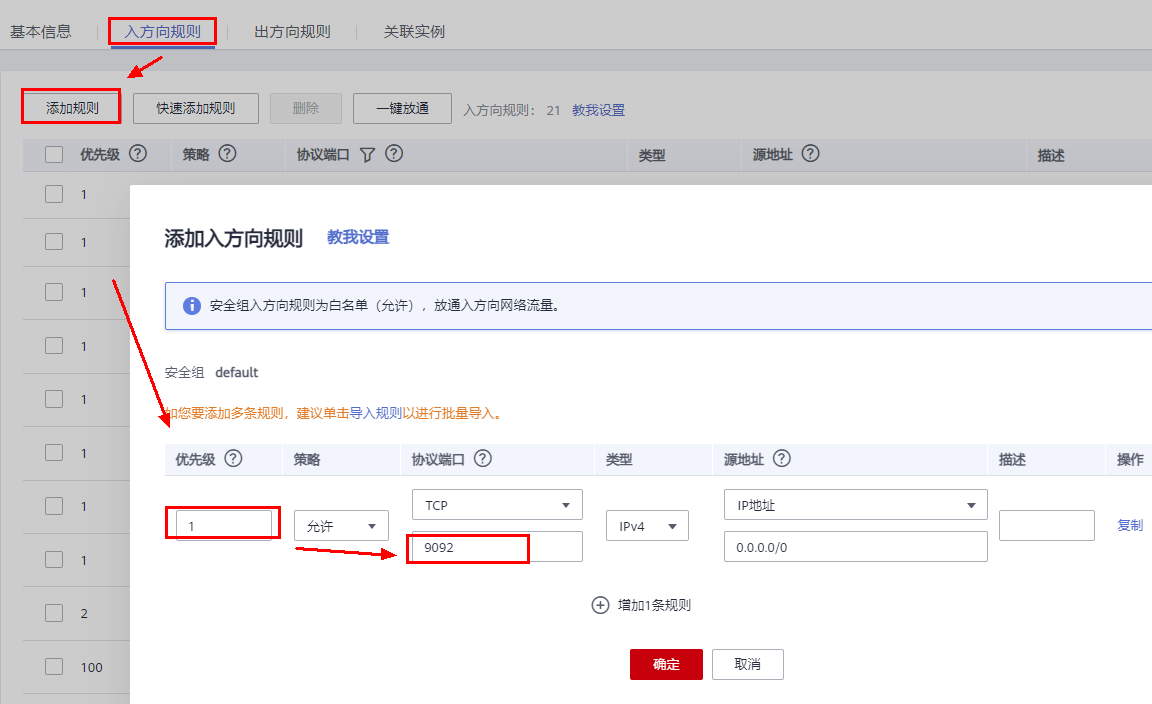
测试DLI与Kafka网络是否连通



输入**kafka\_bootstrap\_servers**地址，测试连通性。



若不可达，则修改安全组，添加对应IP和端口的访问规则。



参考上述方法，再次测试mysql端口连通性，若不可达则再次添加3306端口访问规则。

配置Flink作业的集群信息

进入Flink作业

选择“运行参数”，设置CU数量为2，选择所属队列“queue\_flink”**（必选，为开通DLI服务创建的队列），**其他默认即可。 



运行Flink作业

点击右上角的“启动”。



在启动Flink作业页面点击右下角的“立即启动”。



回到Flink作业界面，状态变为“提交中”。



查看作业运行详情

点击作业管理中的作业名称desc\_order\_count可以进入作业详情页，当作业变为“**运行中**”时可以进行下一步骤操作。（作业提交需要一定的时间，显示“运行中”后就可以查看监控及运行日志）





注：日志中如果出现error字样，则运行错误，请按照对应错误检查相关设置，网络设置、包括安全组是否设置正确，SQL脚本数据库连接是否拼写正确，数据库中是否已经建好相应的数据库和表等问题。

验证数据分析

使用PuTTY登录到MRS的master节点后，进入Python脚本所在目录，再次执行Python脚本生成测试数据：

cd /opt/client/

python autodatagen.py "/tmp/flume\_spooldir/test.txt" 100

回到Flink作业管理界面，点击“作业监控”按钮，进入作业监控页面后，通过指标图表可以看到数据正常处理（各指标图表顺序可以调整）：



登录MySQL后点击数据库的名称或后面的“库管理”进入到库管理界面，可以看到结果表中有数据进来：



点击表名后面的“SQL查询”可以看到表中的数据：



至此，整个数据处理流程完成。

我们新增定时任务，让生成测试数据的Python脚本每隔10秒钟运行一次，在可视化页面上就可以看到统计数据在不断的变化。

# 云资源释放

## 释放数据湖探索服务DLI

1. 删除连接

打开数据湖探索服务页面（如果页面关闭了，可以重新在产品中找到此服务），点击“跨源链接”，点击连接后的删除按钮，删除连接：



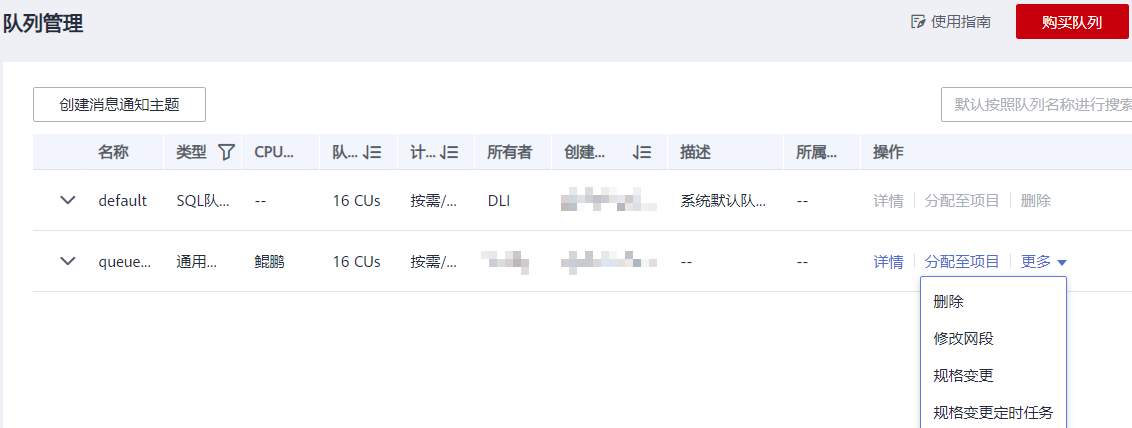
1. 删除作业

点击“作业管理”，点击作业名称后面更多中的删除：



1. 删除队列

点击“队列管理”，点击队列后更多中的删除按钮：



至此，数据湖探索服务（DLI）释放完毕。

## 释放数据管理服务CDM

1. 在控制台的“服务列表”中，点击“迁移”分类下面的“云数据迁移CDM”进入。



1. 删除集群

进入控CDM制台后，点击集群后面“更多”中的“删除”：



弹出对话框点击“确定”，CDM服务资源释放完毕。

## 释放大数据服务MRS

进入MRS控制台，点击集群后的“删除”按钮。

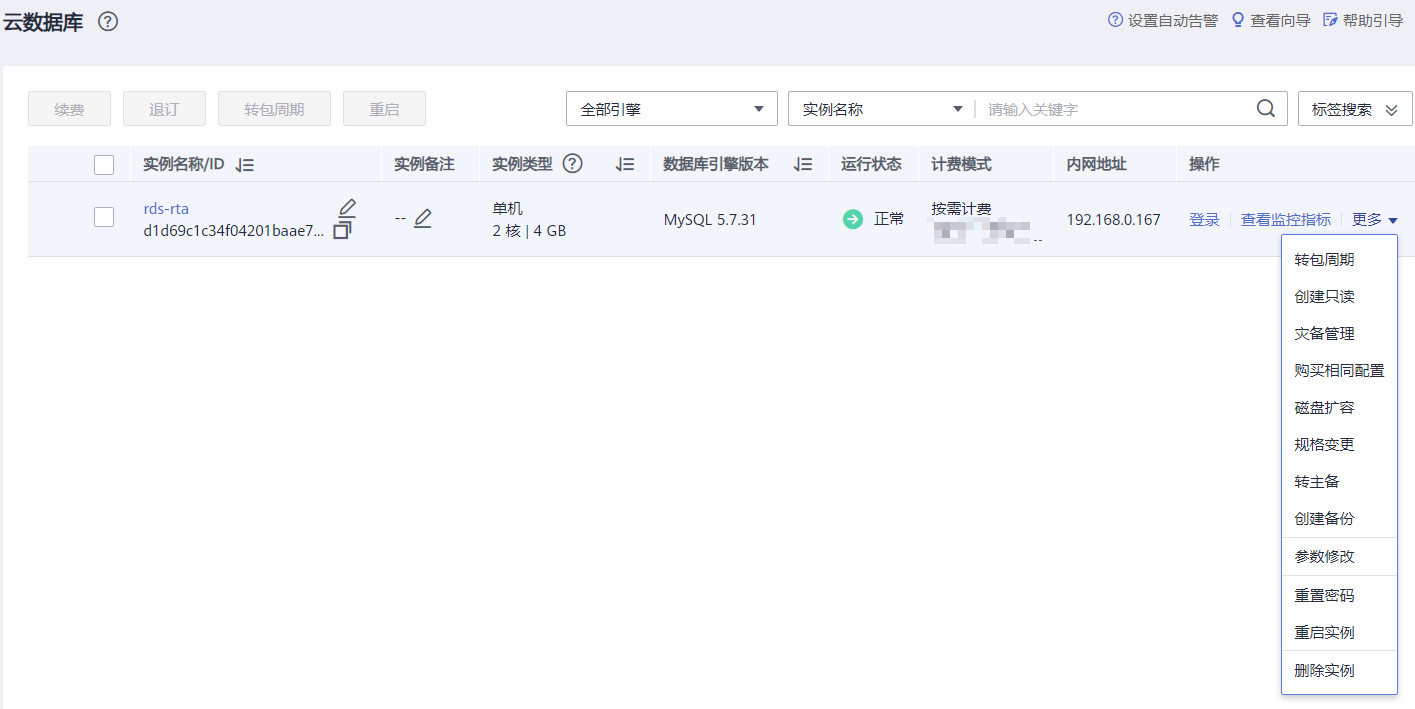


勾选复选框后点击“是”，集群开始删除。



## 释放云数据库RDS

进入云数据库RDS的控制台，在“实例管理”页面点击对应实例后面的“更多”，然后点击弹出菜单中的“删除实例”。



弹出的对话框选择“是”，实例开始删除。



## 释放VPC相关网络资源

1. 进入网络控制台界面

在控制台的服务列表中，点击“网络”分类里的“虚拟私有云VPC”。



进入到虚拟私有云控制台界面。



1. 删除IP

点击“弹性公网IP和带宽”，选择“弹性公网IP”，选中所有要删除的弹性公网IP，点击“更多”里面的“释放”：



弹出对话框选择“是”。



1. 删除安全组规则

点击“访问控制”中的“安全组”，点击要删除的安全组后面“更多”中的“删除”：



弹出对话框点击“是”。

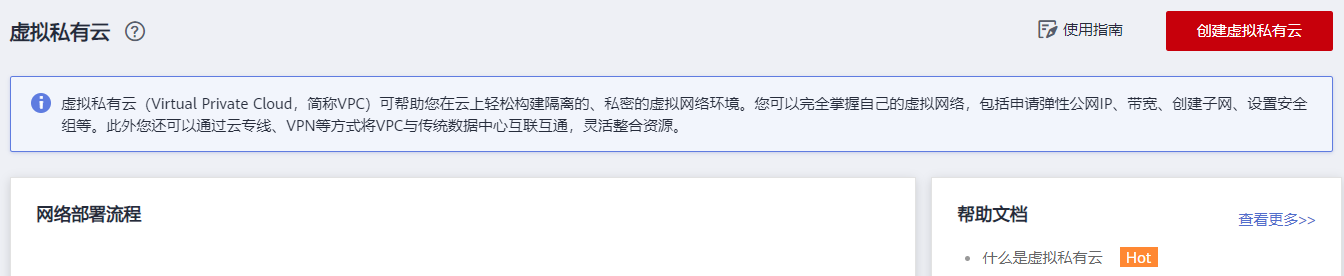


至此，VPC相关资源释放完毕（VPC和子网不需删除）。

## 创建虚拟私有云VPC

1. 创建虚拟私有云

进入虚拟私有云控制台后，点击右上角的“创建虚拟私有云”。



1. 配置基本信息

区域选择“华北-北京四”，名称为“vpc-default”(名称默认或者自定义均可)，点击192网段后面的“选择”。



1. 配置默认子网

子网部分默认即可（名称可以自定义）。



1. 立即创建

点击右下角的“立即创建”。



1. 查看创建的VPC

回到控制台后可以看到创建的VPC。



## 弹性服务器如何绑定弹性IP

1. 进入集群节点管理

点击集群列表中的集群名称，然后再点击节点管理



1. 进入Master所在服务器

点击master\_node\_default\_group下的节点名称， 点击“绑定弹性公网IP”：



选择一个IP，点击确定，完成绑定。



刷新页面，可以看到弹性IP已经绑定成功：



## 如何查看Zookeepr的IP?

1. 登录MRS服务的集群管理页面



1. 使用admin用户登录



1. 查看ZK服务

点击“服务管理”，选择Zookeeper服务，点击“实例”选项，可以看到Zookeeper的业务IP：



## 如何查看Kafka实例Broker的IP?

1. 登录MRS服务的集群管理页面



1. 使用admin用户登录



1. 查看ZK服务

点击“服务管理”，选择Kafka服务，点击“实例”选项，可以看到Kafka的Broker的IP：



## 如何查看服务所在区域的region值？

1. 登录华为云账号

点击账号下的“账号中心”

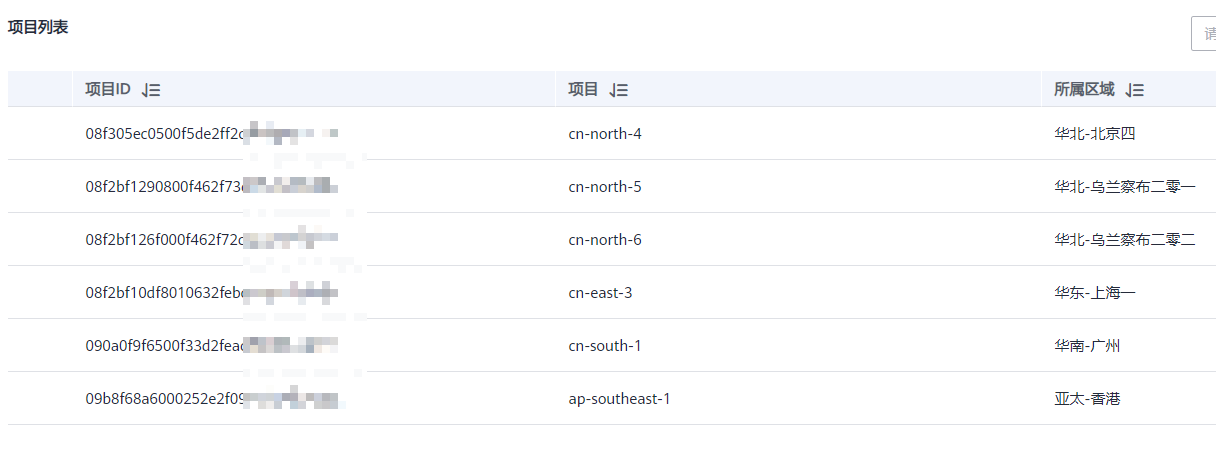


1. 进入“管理我的凭证”



1. 查看区域对应的region

可以看到，华北-北京四对应的项目值就是region的值



## 常用Linux命令

### 命令概述

Linux中的命令格式为：command [options] [arguments] 中括号表示可选的，即有些命令不需要选项也不需要参数，但有的命令在运行时需要多个选项或参数。

* options(选项)：选项是调整命令执行行为的开关，选项的不同决定了命令的显示结果不同。
* agruments(参数)：参数是指命令的作用对象。

### vi/vim

文本编辑器，若文件存在则是编辑，若不存在则是创建并编辑文本。

命令语法：

vim [参数]

参数说明：可编辑的文件名。

命令示例：

* 编辑名为clusterconfig的xml文本：

vim clusterconfig.xml

注：

vim编辑器有以下三种模式：

* 正常模式：其它模式下按Esc或Ctrl+[进入，左下角显示文件名或为空。
* 插入模式：正常模式下按i键进入，左下角显示--INSERT--。
* 可视模式：正常模式下按v键进入，左下角显示--VISUAL--。

退出命令（正常模式下）：

* :wq 保存并退出。
* :q! 强制退出并忽略所有更改。
* :e! 放弃所有修改，并打开原有文件。

### cd

显示当前目录的名称，或切换当前的目录（打开指定目录）。

命令语法：

cd [参数]

参数说明：

* 无参数：切换用户当前目录。
* . ：表示当前目录；
* .. ：表示上一级目录；
* ~ ：表示home目录；
* / ：表示根目录。

命令示例：

* 切换到usr目录下的bin目录中：

cd /usr/bin

* 切换到用户home目录：

cd

* 切换到当前目录(cd后面接一个.)：

cd .

* 切换到当前目录上一级目录(cd后面接两个.)：

cd ..

* 切换到用户home目录：

cd ~

* 切换到根目录下：

cd /

注：切换目录需要理解绝对路径和相对路径这两个概念。

* 绝对路径：在Linux中，绝对路径是从/（即根目录）开始的，例如 /opt/software、/etc/profile, 如果目录以 / 就是绝对目录。
* 相对路径：是以 . 或 .. 开始的目录。 . 表示用户当前操作所在的位置，而 .. 表示上级目录。例如 ./gs\_om 表示当前目录下的文件或者目录。

### mv

文件或目录改名(move (rename) files)或将文件或目录移入其它位置,经常用来备份文件或者目录。

命令语法：

mv [选项] 参数1 参数2

常用选项：

* -b：若需覆盖文件，则覆盖前先行备份。

参数说明：

* 参数1：源文件或目录。
* 参数2：目标文件或目录。

命令示例：

* 将文件python重命名为python.bak：

mv python python.bak

* 将/physical/backup目录下的所有文件和目录移到/data/dbn1目录下：

mv /physical/backup/\* /data/dbn1

### curl

在Linux中curl是一个利用URL规则在命令行下工作的文件传输工具。支持文件的上传和下载，是综合传输工具。

命令语法：

curl [选项] [URL]

常用选项：

* -A/--user-agent <string>：设置用户代理发送给服务器；
* -C/--continue-at <offset>：断点续转；
* -D/--dump-header <file>：把header信息写入到该文件中；
* -e/--referer：来源网址；
* -o/--output：把输出写到该文件中；
* -O/--remote-name：把输出写到该文件中，保留远程文件的文件名；
* -s/--silent：静默模式。不输出任何东西；
* -T/--upload-file <file>：上传文件；
* -u/--user <user[:password]>：设置服务器的用户和密码；
* -x/--proxy <host[:port]>：在给定的端口上使用HTTP代理；
* -#/--progress-bar：进度条显示当前的传送状态。

参数说明：

* URL：指定的文件传输URL地址。

命令示例：

* 将url(<https://mirrors.huaweicloud.com/repository/conf/CentOS-7-anon.repo)的内容保存到/etc/yum.repos.d/CentOS-Base.repo>文件中。

curl -o /etc/yum.repos.d/CentOS-Base.repo <https://mirrors.huaweicloud.com/repository/conf/CentOS-7-anon.repo>

* 如果在传输过程中掉线，可以使用-C的方式进行续传。

curl -C -O <https://mirrors.huaweicloud.com/repository/conf/CentOS-7-anon.repo>

### yum

Shell 前端软件包管理器。基于 RPM 包管理，能够从指定的服务器自动下载 RPM 包并且安装，可以自动处理依赖性关系，并且一次安装所有依赖的软体包，无须繁琐地一次次下载和安装。

命令语法：

yum [options] [command] [package ...]

常用选项：

* -h：查看帮助；
* -y：当安装过程提示选择全部为 "yes"；
* -q：不显示安装的过程。

参数说明：

* command：要进行的操作。
* package：安装的包名。

命令示例：

* 列出所有可更新的软件清单命令：

yum check-update

* 更新所有软件命令：

yum update

* 列出所有可安裝的软件清单命令：

yum list

* 安装指定的软件：

yum install -y libaio-devel flex bison ncurses-devel glibc.devel patch lsb\_release wget python3

### wget

wget是Linux下下载文件的最常用命令。wget支持HTTP,HTTPS和FTP协议,支持自动下载,即可以在用户退出系统后在后台执行,直到下载结束。

命令语法:

wget [选项] [URL]

常用选项:

* -c：接着下载没下载完的文件；
* -b：启动后转入后台执行；
* -P：指定下载目录；
* -O：变更下载文件名；
* --ftp-user --ftp-password：使用FTP用户认证下载。

参数说明:

* 指定的文件下载URL地址。

命令示例：

* 下载openGauss数据库安装文件到当前文件夹：

wget <https://opengauss.obs.cn-south-1.myhuaweicloud.com/1.0.0/x86/openGauss-1.0.0-CentOS-64bit.tar.gz>

* 使用wget断点续传：

wget –c <https://opengauss.obs.cn-south-1.myhuaweicloud.com/1.0.0/x86/openGauss-1.0.0-CentOS-64bit.tar.gz>

### ln

为某一个文件在另外一个位置建立一个同步的链接（软硬链接，不带选项为硬链接）。

当需要在不同的目录，用到相同的文件时，就不需要在每一个需要要的目录下都放一个必须相同的文件，我们只要在某个固定的目录，放上该文件，然后在 其它的目录下用ln命令链接（link）它就可以，不必重复的占用磁盘空间。

命令语法：

ln [选项] 参数1 参数2

常用选项：

* -b --删除，覆盖以前建立的链接；
* -d --允许超级用户制作目录的硬链接；
* -s --软链接(符号链接)。

参数说明：

* 参数1：源文件或目录。
* 参数2：被链接的文件或目录。

命令示例：

* 为python3文件创建软链接/usr/bin/python，如果python3丢失，/usr/bin/python将失效：

ln -s python3 /usr/bin/python

* 为python3创建硬链接/usr/bin/python，python3与/usr/bin/python的各项属性相同：

ln python3 /usr/bin/python

### mkdir

创建指定的名称的目录，要求创建目录的用户在当前目录中具有写权限，并且指定的目录名不能是当前目录中已有的目录。

命令语法：

mkdir [选项] [参数]

常用选项：

* -p --可以是一个路径名称。此时若路径中的某些目录尚不存在,加上此选项后,系统将自动建立好那些尚不存在的目录,即一次可以建立多个目录（递归）；
* -v --每次创建新目录都显示信息；
* -m --设定权限<模式> (类似chmod)，而不是rwxrwxrwx减umask。

参数说明：

* 需要创建的目录。

命令示例：

* 创建一个空目录：

mkdir test

* 递归创建多个目录：

mkdir -p /opt/software/openGauss

* 创建权限为777的目录(目录的权限为rwxrwxrwx)：

mkdir –m 777 test

### chmod

更改文件权限。

命令语法：

chmod [选项] <mode> <file...>

常用选项：

* -R, --以递归的方式对目前目录下的所有文件与子目录进行相同的权限变更。

参数说明：

* mode：权限设定字串，详细格式如下 ：

[ugoa...][[+-=][rwxX]...][,...]，

其中，[ugoa...]：u 表示该档案的拥有者，g 表示与该档案的拥有者属于同一个群体(group)者，o 表示其他以外的人，a 表示所有（包含上面三者）；[+-=]：+ 表示增加权限，- 表示取消权限，= 表示唯一设定权限；[rwxX]：r 表示可读取，w 表示可写入，x 表示可执行，X表示只有当该档案是个子目录或者该档案已经被设定过为可执行。

* file：文件列表（单个或者多个文件、文件夹）。

命令示例：

* 设置所有用户可读取文件 cluterconfig.xml：

chmod ugo+r cluterconfig.xml

或

chmod a+r cluterconfig.xml

* 设置当前目录下的所有档案与子目录皆设为任何人可读写：

chmod -R a+rw \*

数字权限使用格式：

* 这种使用方式中，规定数字4、2和1表示读、写、执行权限，即r=4,w=2,x=1。
* 例：rwx = 7（4+2+1）；rw = 6（4+2）；r-x = 5 （4+0+1）；r-- = 4（4+0+0）；--x = 1（0+0+1）；

每个文件都可以针对三个粒度，设置不同的rwx(读写执行)权限。即我们可以用用三个8进制数字分别表示 拥有者 、群组 、其它组( u、 g 、o)的权限详情，并用chmod直接加三个8进制数字的方式直接改变文件权限。语法格式为 ：

chmod <abc> file...

其中，a,b,c各为一个数字，分别代表User、Group、及Other的权限，相当于简化版的chmod u=权限,g=权限,o=权限 file...，而此处的权限将用8进制的数字来表示User、Group、及Other的读、写、执行权限。

命令示例：

* 赋予cluterconfig.xml文件可读可写可执行权限（所有权限）：

chmod 777 cluterconfig.xml

* 赋予/opt/software/openGauss目录下所有文件及其子目录 用户所有权限组可读可执行权限，其他用户可读可执行权限：

chmod R 755 /opt/software/openGauss

### chown

利用 chown 将指定文件的拥有者改为指定的用户或组，用户可以是用户名或者用户ID；组可以是组名或者组ID；文件是以空格分开的要改变权限的文件列表，支持通配符。只有系统管理者(root)才有这样的权限。使用权限 : **root**。

命令语法：

chown [选项] user[:group] file...

常用选项:

* -c : 显示更改的部分的信息；
* -f : 忽略错误信息；
* -R : 处理指定目录以及其子目录下的所有文件。

参数说明

* user : 新的文件拥有者的使用者 ID。
* group : 新的文件拥有者的使用者组(group)。
* flie：文件。

命令示例：

* 将文件 file1.txt 的拥有者设为omm，群体的使用者dbgrp:

chown omm:dbgrp /opt/software/openGauss/clusterconfig.xml

* 将目前目录下的所有文件与子目录的拥有者皆设为omm，群体的使用者dbgrp:

chown -R omm:dbgrp \*

### ls

列出文件和目录的内容。

命令语法：

ls [选项] [参数]

常用选项：

* -l --以长格式显示，列出文件的详细信息，如创建者，创建时间，文件的读写权限列表等等；
* -a --列出文件下所有的文件，包括以"."和".."开头的隐藏文件 （Linux下文件隐藏文件是以 .开头的，如果存在 .. 代表存在着父目录）；
* -d --列出目录本身而非目录内的文件，通常要与-l一起使用；
* -R --同时列出所有子目录层，与-l相似，只是不显示出文件的所有者，相当于编程中的“递归”实现；
* -t --按照时间进行文件的排序，Time（时间）；
* -s --在每个文件的后面打印出文件的大小，size（大小）；
* -S --以文件的大小进行排序。

参数说明：

* 目录或文件。

命令示例：

* 以长格式列出当前目录中的文件及目录：

ls -l

### cp

复制文件或者目录。

命令语法：

cp [选项] 参数1 参数2

常用选项：

* -f --如果目标文件无法打开则将其移除并重试(当 -n 选项存在时则不需再选此项)；
* -n --不要覆盖已存在的文件(使前面的 -i 选项失效)；
* -I --覆盖前询问(使前面的 -n 选项失效)；
* -p --保持指定的属性(默认：模式,所有权,时间戳)，如果可能保持附加属性：环境、链接、xattr 等；
* -R,-r --复制目录及目录内的所有项目。

参数说明：

* 参数1：源文件。
* 参数2：目标文件。

命令示例：

* 将home目录中的abc文件复制到opt目录下：

cp /home/abc /opt

注：目标文件存在时，会询问是否覆盖。这是因为cp是cp -i的别名。目标文件存在时，即使加了-f标志，也还会询问是否覆盖。

### rm

删除一个目录中的一个或多个文件或目录，它也可以将某个目录及其下的所有文件及子目录均删除。对于链接文件，只是删除了链接，原有文件均保持不变。

rm是一个危险的命令，使用的时候要特别当心，否则整个系统就会毁在这个命令（比如在/（根目录）下执行rm \* rf）。所以，我们在执行rm之前最好先确认一下在哪个目录，到底要删除什么东西，操作时保持高度清醒的头脑。

命令语法：

rm [选项] 文件

常用选项：

* -f --忽略不存在的文件，从不给出提示；
* -r --指示rm将参数中列出的全部目录和子目录均递归地删除。

参数说明：

* 需要删除的文件或目录。

命令示例：

* 删除文件：

rm qwe

注：输入rm qwe命令后，系统会询问是否删除，输入y后就会删除文件，不想删除文件则输入n。

* 强制删除某个文件：

rm-rf clusterconfig.log

### cat

连接文件并在标准输出上输出。这个命令常用来显示文件内容，或者将几个文件连接起来显示，或者从标准输入读取内容并显示，它常与重定向符号配合使用。

命令语法：

cat [选项] [参数]

常用选项：

* -E --在每行结束显示$；
* -n –由1开始对给所有输出行编号；
* -b 或 --number-nonblank：和 -n 相似，只不过对于空白行不编号；
* -v --使用 ^ 和 M- 符号，除了 LFD 和 TAB 之外。

参数说明：

* 可操作的文件名。

命令示例:

* 显示testfile文件的内容：

cat textfile

* 把 textfile1 和 textfile2 的文档内容加上行号（空白行不加）之后将内容追加到 textfile3 文档里：

cat -b textfile1 textfile2 >> textfile3

* 向/etc/profile中追加内容（输入EOF表示结束追加）：

cat >>/etc/profile<<EOF

>export LD\_LIBRARY\_PATH=$packagePath/script/gspylib/clib:$LD\_LIBRARY\_PATH

>EOF

注：

* EOF是end of file的缩写，表示"文字流"（stream）的结尾。"文字流"可以是文件（file），也可以是标准输入（stdin）。在Linux系统之中，EOF是当系统读取到文件结尾，所返回的一个信号值（也就是-1）。

### unzip

解压、查看、测试zip文档。

命令语法：

unzip [选项] 文件名.zip

常用选项：

* -a 对文本文件进行必要的字符转换；
* -c 将解压缩的结果显示到屏幕上，并对字符做适当的转换；
* -f 更新现有的文件；
* -l 显示压缩文件内所包含的内容；
* -o 不询问用户，unzip执行后直接覆盖原文件；
* -t 检查压缩文件是否正确；
* -v 执行是时显示详细的信息；
* -z 仅显示压缩文件的备注文字。

参数说明：

* 可操作的zip文件名。

命令示例:

* 解压一个文件：

unzip demo.zip

* 查看压缩文件的目录信息，而不解压：

unzip -v demo.zip

* 查看压缩文件的内容：

unzip -l demo.zip

### source

使shell读入指定的文件并依次执行文件中的所有语句。shell通常用“.”代替，被称为点命令。通常用于重新执行刚修改的初始化文件，使之立即生效，而无需注销并重新登录。

命令语法：

source 文件名 或者 . 文件名

注意：

* 文件名对应的必须是可执行的脚本文件；
* 点和文件名之间要保留一个空格。

参数说明：

* 可执行的脚本文件名。

命令示例:

* 执行一个文件：

source env\_pro

### more

此命令类似于cat命令，但显示时会以页的形式进行翻屏方便阅读。按空白键（space）向下翻一页，按 b 键往回（back）翻一页。

命令语法：

more 文件名

常用选项：

* -num 一次要显示的行数；
* -f 以文件实际的行数计算行数，而不是自动换行过后的行数；
* -p 先清除屏幕再显示每一页内容，而不是以卷动的方式显示；
* -s 遇到有连续两行以上的空白行时替换为一行的空白行；
* +num 从第num行开始显示。

参数说明：

* 要翻阅的文件名。

常用命令：

* Enter 向下一行；
* = 输出当前行号；
* Space 向下滚动一屏；
* b 向上滚动一屏；
* q 退出more。

命令示例:

* 从第 10 行开始显示 myfile 内容：

more +10 myfile

### tar

用于备份文件，可以建立或者还原备份文件。扩展名为.tar的文件通常称为tar包。

命令语法：

tar 选项 参数 文件

常用选项：

* -A或--catenate 新增文件到已存在的备份文件；
* -c或--create 创建新的备份文件；
* -C<目录> 解压到指定的目录；
* -f<备份文件> 指定备份文件；
* -r新增文件到已存在的备份文件的结尾部分；
* -t列出备份文件的内容；
* -v显示指令执行过程；
* -x从备份文件中还原文件；
* -z通过gzip指令处理备份文件。

参数说明：

* 要操作的文件名。

命令示例:

* 列出压缩文件内容：

tar -tzvf myfile.tar.gz

* 压缩文件a.txt为myfile.tar.gz：

tar -czvf test.tar.gz a.c

* 解压文件myfile.tar.gz：

tar -xzvf test.tar.gz a.c

## 实验FAQ

1、数据采集成功，但是在mysql中没有数据结果

1. 确认流作业里面的mysql参数是否正确，和数据库里的表&列名称是否匹配，集群是否选择正确；
2. agent.yml文件中的ak、sk、projectid、DISStream、filepattern是否正确；
3. 作业是否正常提交，agent运行窗口是否上传数据正常，有无报错；
4. 网络是否正常，确认对等连接已连接正常，安全组是否正确设置；
5. DIS中显示是否正常；
6. 查看实时流计算中作业管理下的运行日志，有无报错。

2、找不到服务

本实验很多华为云产品，会在各个页面间跳转，如果不小心关闭了某个服务页面，找不到原来的服务，可以通过首页—产品，再次找到此服务，进入到它的控制台页面操作。

五：实验总结(10’)