

# 1-创建 MRS 分析集群&使用 MRS 客户端

## 1. 任务介绍

本次任务将介绍如何进行 MRS 服务集群的创建和 MRS 客户端的基本使用，完成驾驶行为分析。

## 2. 任务执行

### 2.1 申请虚拟私有云（已有虚拟私有云可跳过本步骤）

2.1.1 登录[华为云控制台](#)，选择“网络 > [虚拟私有云](#)”，确认左上角的区域选择为“北京一”。



2.1.2 在页面右上角中选择“创建虚拟私有云”。

2.1.3 在新打开的页面中填写虚拟私有云的基本信息，按照以下方式填写：

参数	值
区域	华北-北京一
名称	vpc-mrs-demo
网段	默认值
标签	默认值
可用分区	默认值
子网名称	subnet-mrs-demo
子网网段	默认值
高级配置	默认配置

配置完成以后如下图所示：

基本信息

区域

华北-北京一

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，可以降低网络时延、提高访问速度。

\* 名称

vpc-mrs-demo

\* 网段

192 · 168 · 0 · 0 / 16

建议使用网段：10.0.0.0/8~24，172.16.0.0/12~24，192.168.0.0/16~24

标签

标签键

标签值

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](#)

您还可以添加10个标签。

子网配置

默认子网

可用区 ?

可用区1

可用区2

可用区3

\* 名称

subnet-mrs-demo

\* 子网网段

192 · 168 · 0 · 0 / 24

可用IP数:250 子网创建完成后，子网网段无法修改

高级配置

默认配置

自定义配置

2.1.4 检查当前配置后单击“立即创建”。

2.2 购买 MRS 集群

2.2.1 登录[华为云控制台](#)，选择“EI 企业智能 > [MapReduce 服务](#)”。

2.2.2 在页面中选择“购买集群”，进入“集群配置”页面。

2.2.3 在新打开的页面中填写集群的基本信息（请按照文档的要求进行配置，否则代金券不足以支撑到实验结束）：

参数	值
计费模式	按需计费
当前区域	默认值：北京一
可用分区	默认值：可用区 2
集群名称	21days_username（以自己的华为云用户名作为集群名称）
集群版本	1.7.2

Kerberos 认证	默认值：关闭
集群类型	默认值：分析集群
组件选择	全选
虚拟私有云	选择之前创建的 VPC：vpc-mrs-demo
子网	选择之前创建 VPC 对应的子网：sunbet-mrs-demo
安全组	默认值：自动创建
集群高可用	关闭
集群节点	Master 的实例规格：通用计算增强型 C3 4 核 16GB
	Core 的实例规格：通用计算增强型 C3 4 核 16GB
	Core 的实例数量：1
登录方式	默认值：密码
高级配置	默认值：暂不配置

**注意：如果集群节点的规格置售罄请按照如下方法尝试：**

- 1. 优先切换可用区，查看其它可用区该规格是否仍有资源**
- 2. 如果其它可用区没有 C3 规格，请选择 S3 4 核 16GB 规格。**

配置完成以后部分信息如下图所示，最终费用为 3.31 元/小时即代表配置正确：

\* 计费模式

包年/包月

按需计费

\* 当前区域

华北-北京一

\* 可用分区

可用区1

可用区2

\* 集群名称

21days\_username

\* 集群版本

MRS 1.7.2

\* Kerberos认证

☐

未开启认证，存在安全风险。了解更多

\* 集群类型

分析集群

流式集群

\* 组件选择

组件名	版本	描述	<input checked="" type="checkbox"/>
Hadoop	2.8.3	针对大数据集的分布式数据处理...	<input checked="" type="checkbox"/>
Spark	2.2.1	快速、通用的大数据处理引擎	<input checked="" type="checkbox"/>
HBase	1.3.1	可扩展、分布式数据库，支持存...	<input checked="" type="checkbox"/>
Hive	1.2.1	提供数据汇聚和即席查询的数据...	<input checked="" type="checkbox"/>
Hue	3.11.0	提供hadoop UI能力，让用户通过...	<input checked="" type="checkbox"/>
Loader	2.0.0	Loader是基于开源Sqoop 1.99.7...	<input checked="" type="checkbox"/>

\* 虚拟私有云

vpc-mrs-demo

查看虚拟私有云

\* 子网

subnet-mrs-demo(1...

\* 安全组

自动创建

管理安全组

\* 集群高可用

☐

\* 集群节点

类型	实例规格	实例数	数据盘	弹性伸缩	操作
Master	4 核 16 GB   c3.xlarg...	1	普通IO	200 GB x 1	--
Core	4 核 16 GB   c3.xlarg...	1	普通IO	100 GB x 1	--

配置费用

¥ 3.31/小时

参考价格，具体扣费请以账单为准 了解详情

立即购买

2.2.4 单击右下角的“立即申请”，在下一个页面中确认集群配置和服务协议后，单击“提交申请”。

2.2.5 页面跳转到“[集群列表](#)”页面后，可以看到申请的集群正在启动中，集群创建需要一段时间，可先执行下步操作。

### 2.3 申请弹性 IP

2.3.1 登录[虚拟私有云地址](#)，点击左侧“弹性公网 IP”页签，进入弹性公网 IP 页面。点击右上角的“购买弹性公网 IP”进入到申请页面。

网络控制台

弹性公网IP

弹性公网IP

IP v4公网中，按需您立即体验：IP v4 IP  
公网期间IPv6转换功能免费，带宽正常收费。

您还可以购买20个弹性公网IP。

解绑 释放 续费 按小时/包月

所有状态 弹性公网IP 搜索 帮助 刷新 分享

弹性公网IP ID	状态	类型	带宽	带宽详情	已绑定实例	计费模式	操作

2.3.2 在打开的页面中填写弹性 IP 的基本信息：

参数	值
----	---

计费模式	按需计费
区域	华北-北京一
类型	默认值：全动态 BGP
带宽类型	默认值：独享带宽
计费方式	按流量收费（只有从华为云出口的流量才计费，例如：上传数据到华为云是不收费的，从华为云下载数据是计费的）
带宽大小	默认值：5 Mbit/s
带宽名称	mrs-demo
标签	默认值：（空）
购买量	1

配置完成后部分信息如图：

计费模式

包年/包月

按需计费

区域

华北-北京一

类型

全动态BGP

静态BGP

带宽类型

独享带宽

共享带宽

计费方式

按带宽计费

按流量计费

带宽大小(Mbit/s)

1

100

200

300

5

带宽名称

mrs-demo

标签

请输入标签键

请输入标签值

购买量

-

1

+

弹性公网IP费用 ¥0.02/小时 + 公网流量费用 ¥0.80/GB

参考价格，具体扣费请以账单为准。 [了解计费详情](#)

立即购买

2.3.3 单击右下角的“立即申请”，在下一个页面中确认资源详情后，单击“提交”。

## 2.4 绑定弹性 IP

2.4.1 返回[集群列表](#)，点击创建的集群名称，进入到集群管理页面。

集群列表

• 现有集群

• 历史集群

操作日志

帮助

您还剩余8台弹性云服务器、245个：

名称

21days\_username

2.4.2 从节点列表中找到“类型”为“Master1”的节点，点击名称进入到云服务器控制台页面。

节点信息

作业管理

文件管理

告警列表

调整集群

您当前已有1个Core节点，当前可用资源最多可以创建3个节点。[申请扩大配额](#)

名称	状态	类型
d16d60dd-092e-4c75-93e5-6df35323e1e9_node_core_lcPEV	运行中	Core
d16d60dd-092e-4c75-93e5-6df35323e1e9_node_master1_IIXTX	运行中	Master1

2.4.3 点击“弹性 IP”标签页，然后点击“绑定弹性 IP”按钮，选择之前创建的弹性 IP，点击确定即可完成绑定。

云服务器控制台

总览

弹性云服务器

云服务器备份

裸金属服务器

弹性GPU云服务

云硬盘

专属存储

云硬盘备份

镜像服务

弹性伸缩

弹性负载均衡

密钥对

云服务器组

弹性云服务器 > d16d60dd-092e-4c75-93e5-6df35323e1e9\_node\_master1\_IIXTX

名称: d16d60dd-092e-4c75-93e5-6df35323e1e9\_node\_master1\_IIXTX

状态: 运行中

ID: 3c5796e8-8753-4dda-9012-10798d3c3133

磁盘: 2个

可用区: 可用区2

计费模式: 按需

许可类型: 无

代理名称: LogCollectio  
n

云硬盘 网卡 安全组 弹性IP 监控

绑定弹性IP 查看弹性IP

## 2.5 登录服务器并上传数据

2.5.1 返回[集群列表](#)，等到集群状态为“运行中”再执行下面的操作。

2.5.2 使用登录工具（如 PuTTY 等）通过 ssh 的方式连接弹性 IP，输入 root 密码后即可登录集群的管理节点。

2.5.3 执行以下命令加载环境变量并进行数据下载和解压。

```
source /opt/client/bigdata_env
cd /opt
wget https://obs-mrsdevcloud-public.obs-website.cn-north-1.myhwclouds.com/detail-records.zip
unzip /opt/detail-records.zip
```

## 2.6 使用 HDFS 客户端将数据上传到 HDFS

2.6.1 执行如下命令完成数据上传

```
source /opt/client/bigdata_env
hdfs dfs -mkdir /user/hdfs-examples/
hdfs dfs -put /opt/detail-records /user/hdfs-examples/
```

附：执行如下命令可以查看 hdfs 命令的使用帮助，或者到 [HDFS Shell 命令使用帮助文档](#) 进行学习。

```
hdfs dfs -help
```

## 2.7 使用 Spark 客户端建表并查询

### 2.7.1 执行如下命令进入 spark 客户端

```
source /opt/client/bigdata_env
spark-beeline
```

2.7.2 然后输入如下 sql 语句来完成建表的操作（以下语句为一条 sql，为了方便查看进行了分行处理，可在文档将 username 修改后全部复制到 spark 客户端中，然后回车执行）。

```
create external table if not exists username (  
    driverID String,carNumber String,latitude String,longitude String,speed  
String,direction  
    String,siteName String,time String,isRapidlySpeedup String,isRapidlySlowdown  
    String,isNeutralSlide String,isNeutralSlideFinished String,neutralSlideTime  
    String,isOverspeed String,isOverspeedFinished String,overspeedTime  
    String,isFatigueDriving String,isHthrottleStop String,isOilLeak String)  
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
STORED AS TEXTFILE  
LOCATION '/user/hdfs-examples/detail-records';
```

2.7.3 使用如下 sql 完成统计查询，其中标红部分的表名以实际为准（说明同上）。

```
select  
    driverID,  
    carNumber,  
    sum(isRapidlySpeedup) as rapidlySpeedupTimes,  
    sum(isRapidlySlowdown) as rapidlySlowdownTimes,  
    sum(isNeutralSlide) as neutralSlideTimes,  
    sum(neutralSlideTime) as neutralSlideTimeTotal,  
    sum(isOverspeed) as overspeedTimes,  
    sum(overspeedTime) as overspeedTimeTotal,  
    sum(isFatigueDriving) as fatigueDrivingTimes,  
    sum(isHthrottleStop) as hthrottleStopTimes,  
    sum(isOilLeak) as oilLeakTimes  
from  
    username
```



```
where
    time >= '2017-01-01 00:00:00'
    and time <= '2017-02-01 00:00:00'
    and (isRapidlySpeedup > 0
    OR isRapidlySlowdown > 0
    OR isNeutralSlide > 0
    OR isNeutralSlideFinished > 0
    OR isOverspeed > 0
    OR isOverspeedFinished > 0
    OR isFatigueDriving > 0
    OR isHthrottleStop > 0
    OR isOilLeak > 0)
group by
    driverID,
    carNumber
order by
    rapidlySpeedupTimes desc,
    rapidlySlowdownTimes desc,
    neutralSlideTimes desc,
    neutralSlideTimeTotal desc,
    overspeedTimes desc,
    overspeedTimeTotal desc,
    fatigueDrivingTimes desc,
    hthrottleStopTimes desc,
    oilLeakTimes desc;
```

2.7.4 最终程序执行如下图：

```

0: jdbc:hive2://ha-cluster/default> select
0: jdbc:hive2://ha-cluster/default> driverID,
0: jdbc:hive2://ha-cluster/default> carNumber,
0: jdbc:hive2://ha-cluster/default> sum(isRapidlySpeedup) as rapidlySpeedupTimes,
0: jdbc:hive2://ha-cluster/default> sum(isRapidlySlowdown) as rapidlySlowdownTimes,
0: jdbc:hive2://ha-cluster/default> sum(isNeutralSlide) as neutralSlideTimes,
0: jdbc:hive2://ha-cluster/default> sum(neutralSlideTime) as neutralSlideTimeTotal,
0: jdbc:hive2://ha-cluster/default> sum(isOverspeed) as overspeedTimes,
0: jdbc:hive2://ha-cluster/default> sum(overspeedTime) as overspeedTimeTotal,
0: jdbc:hive2://ha-cluster/default> sum(isFatigueDriving) as fatigueDrivingTimes,
0: jdbc:hive2://ha-cluster/default> sum(isHtrrottleStop) as htrrottleStopTimes,
0: jdbc:hive2://ha-cluster/default> sum(isOilLeak) as oilLeakTimes
0: jdbc:hive2://ha-cluster/default> from
0: jdbc:hive2://ha-cluster/default>   username
0: jdbc:hive2://ha-cluster/default> where
0: jdbc:hive2://ha-cluster/default>   time >= '2017-01-01 00:00:00'
0: jdbc:hive2://ha-cluster/default>   and time <= '2017-02-01 00:00:00'
0: jdbc:hive2://ha-cluster/default>   and (isRapidlySpeedup > 0
0: jdbc:hive2://ha-cluster/default>   OR isRapidlySlowdown > 0
0: jdbc:hive2://ha-cluster/default>   OR isNeutralSlide > 0
0: jdbc:hive2://ha-cluster/default>   OR isNeutralSlideFinished > 0
0: jdbc:hive2://ha-cluster/default>   OR isOverspeed > 0
0: jdbc:hive2://ha-cluster/default>   OR isOverspeedFinished > 0
0: jdbc:hive2://ha-cluster/default>   OR isFatigueDriving > 0
0: jdbc:hive2://ha-cluster/default>   OR isHtrrottleStop > 0
0: jdbc:hive2://ha-cluster/default>   OR isOilLeak > 0)
0: jdbc:hive2://ha-cluster/default> group by
0: jdbc:hive2://ha-cluster/default>   driverID,
0: jdbc:hive2://ha-cluster/default>   carNumber
0: jdbc:hive2://ha-cluster/default> order by
0: jdbc:hive2://ha-cluster/default>   rapidlySpeedupTimes desc,
0: jdbc:hive2://ha-cluster/default>   rapidlySlowdownTimes desc,
0: jdbc:hive2://ha-cluster/default>   neutralSlideTimes desc,
0: jdbc:hive2://ha-cluster/default>   neutralSlideTimeTotal desc,
0: jdbc:hive2://ha-cluster/default>   overspeedTimes desc,
0: jdbc:hive2://ha-cluster/default>   overspeedTimeTotal desc,
0: jdbc:hive2://ha-cluster/default>   fatigueDrivingTimes desc,
0: jdbc:hive2://ha-cluster/default>   htrrottleStopTimes desc,
0: jdbc:hive2://ha-cluster/default>   oilLeakTimes desc;

```

driverID	carNumber	rapidlySpeedupTimes	rapidlySlowdownTimes	neutralSlideTimes	neutralSlideTimeTotal	overspeedTimes	overspeedTimeTotal	fatigueDrivingTimes	htrrottleStopTimes	oilLeakTimes
hanhui10000002	4#A21419	461.0	444.0	327.0	2844.0	3349.0	31813.0	3997.0	433.0	371.0
panxian10000005	4#A0C42C	395.0	434.0	336.0	2930.0	3531.0	35946.0	4367.0	417.0	441.0
shenxian10000004	4#A01759	372.0	355.0	297.0	2810.0	3125.0	31494.0	3767.0	383.0	365.0
zouan10000007	4#A59M63	360.0	385.0	315.0	2997.0	3181.0	31248.0	3594.0	399.0	385.0
likun10000003	4#A09626	341.0	354.0	291.0	2643.0	2644.0	26726.0	3552.0	347.0	376.0
zhangpeng10000008	4#A21119	340.0	344.0	272.0	2894.0	2793.0	25479.0	3274.0	284.0	337.0
haowei10000008	4#A7096B	321.0	314.0	255.0	2659.0	2639.0	25522.0	3204.0	312.0	318.0
xiexia10000003	4#A2112P	264.0	261.0	246.0	2525.0	2334.0	23434.0	2726.0	314.0	253.0
xiexi10000006	4#A6Q111	255.0	310.0	254.0	2074.0	2535.0	23942.0	2593.0	312.0	279.0
duxu10000009	4#A175H8	238.0	284.0	247.0	2632.0	2301.0	22338.0	2814.0	264.0	248.0

### 3. 打卡任务

完成步骤 2.7.4，并将最终结果截图，需要截出以华为云用户名命名的表名字段。

### 4. 关闭资源

在打卡完成后，可选择终止资源停止扣费。

1. 关闭 MRS 资源。进入 [MapReduce 服务](#)，点击集群右边的“终止”按钮关闭 MRS 服务。

可用分区	操作
可用区2	终止

2. 关闭弹性公网 IP 资源。登录[虚拟私有云地址](#)，点击左侧“弹性公网 IP”页签，进入弹性公网 IP 页面。点击操作中的“更多->释放”按钮进行释放弹性 IP 资源。