

## 1. 项目介绍

### 1.1 项目背景

本次项目使用爬虫程序从天气网站上获取全国各省、市、区县近几年的天气数据（本项目以北京市历史数据为例：<http://www.tianqihoubao.com/lishi/index.htm>），然后对数据进行实时分发、收集和统计分析（例如统计某个区县近几年各类天气情况的分布或者了解某年某月哪个地区温度最高），来引导用户熟悉华为云的 MRS 服务的使用，最后通过 BI 工具进行图表展示。

### 1.2 数据介绍

部分数据展示如图，其中每行数据代表一次天气记录。

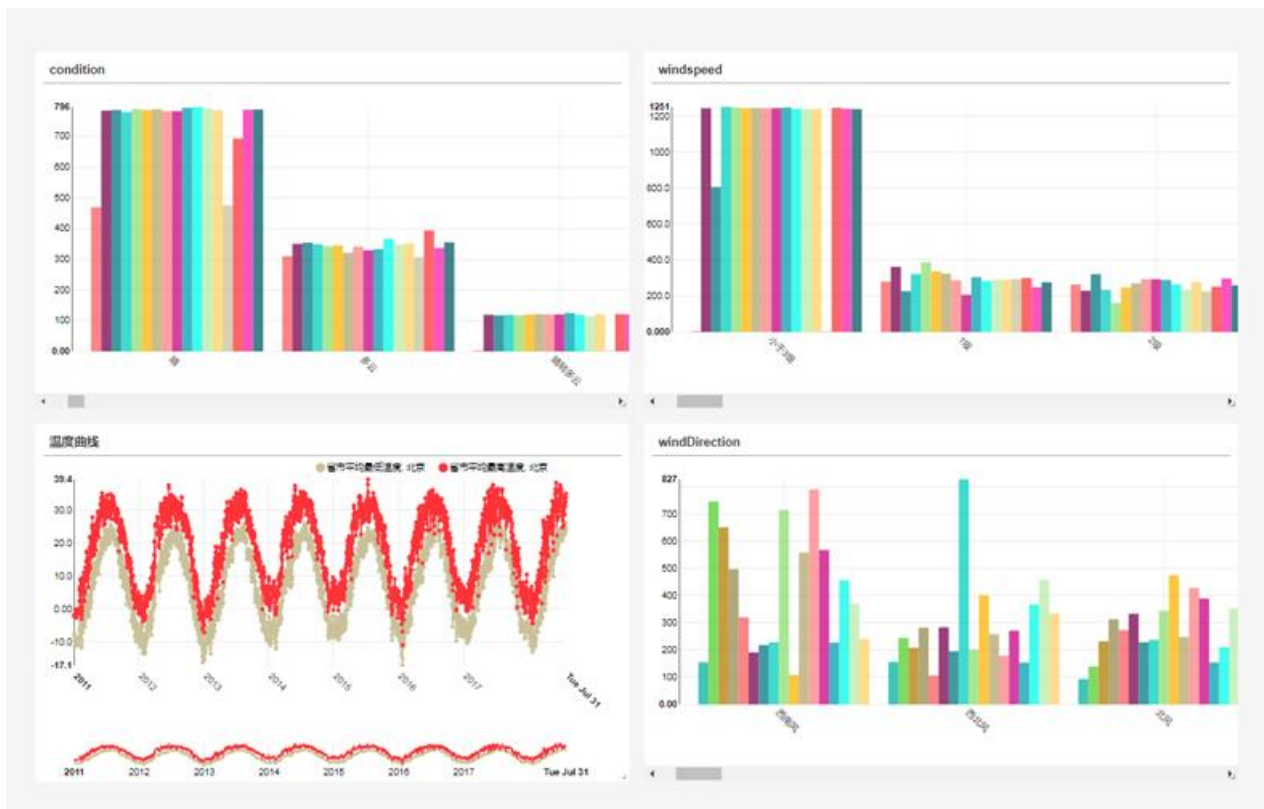
地区ID	省	市	地区	时间	天气	最高温度	最低温度	风力
yanqing	北京	北京	延庆	2017/6/1	晴/阵雨	29	13	南风 3-4级 / 东风 ≤3级
yanqing	北京	北京	延庆	2017/6/2	阴/多云	21	10	东风 3-4级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/3	晴/晴	25	13	南风 3-4级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/4	晴/多云	28	15	西南风 3-4级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/5	多云/小雨	26	12	南风 3-4级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/6	小雨/多云	17	11	南风 ≤3级 / 西北风 ≤3级
yanqing	北京	北京	延庆	2017/6/7	多云/多云	28	14	南风 ≤3级 / 南风 ≤3级
yanqing	北京	北京	延庆	2017/6/8	阵雨/晴	32	17	南风 ≤3级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/9	晴/多云	34	17	南风 3-4级 / 南风 ≤3级
yanqing	北京	北京	延庆	2017/6/10	多云/阵雨	27	14	东南风 ≤3级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/11	阵雨/阵雨	29	14	南风 ≤3级 / 东风 ≤3级
yanqing	北京	北京	延庆	2017/6/12	阵雨/阵雨	24	14	南风 ≤3级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/13	雷阵雨/晴	26	13	南风 ≤3级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/14	晴/晴	34	17	南风 ≤3级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/15	晴/晴	35	20	南风 ≤3级 / 北风 ≤3级
yanqing	北京	北京	延庆	2017/6/16	多云/阴	35	20	东南风 3-4级 / 东南风 3-4级

数据名称和对应的变量名

地区ID	省	市	地区	时间	天气	最高温度	最低温度	风力
id	province	city	zone	time	weather	maxTemperature	minTemperature	windPower

### 1.3 结果展示

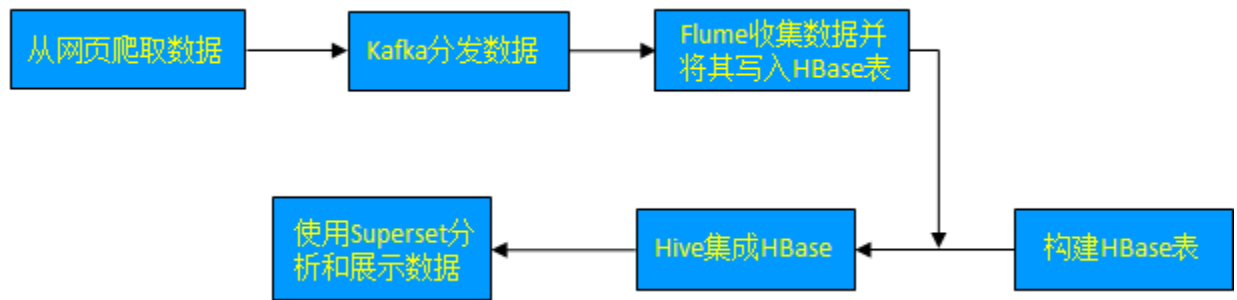
最终可以在 BI 工具 superset 中对数据进行各种统计分析和图表展示。



## 2. 解决方案流程介绍

### 2.1 总体执行流程

- 首先需要构建 HBase 表。
- 从网页上爬取数据。
- 使用 kafka 对数据进行分发。
- 利用 Flume 收集数据并将其写入 HBase 表。
- 构建 Hive 表并与 HBase 表进行关联。
- 使用 Superset 统计和展示 Hive 表中的数据。



### 3. 创建 MRS 分析集群

#### 3.1 申请虚拟私有云（已有虚拟私有云可跳过本步骤）

3.1.1 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“华北-北京一”。



3.1.2 在页面右上角中选择“创建虚拟私有云”。

3.1.3 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
可用区	可用区 2
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

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

您还可以添加10个标签。

子网配置

默认子网

可用区

可用区2

可用区1

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

立即创建

3.1.4 检查当前配置后单击“立即创建”。

## 3.2 购买 MRS 分析集群（已购买分析集群可跳过此步骤）

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

3.2.2 在页面中选择“购买集群”，进入“集群配置”页面。

3.2.3 在新打开的页面中填写集群的基本信息（“集群高可用”和“集群节点”配置在本次实践将采用低成本配置）：

参数	值
计费模式	按需计费
当前区域	默认值：华北-北京一
可用分区	默认值：可用区 2
集群名称	mrs_demo_analysis
集群版本	1.7.2

2018-11-30

华为保密信息,未经授权禁止扩散

第 4 页，共 26 页

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	暂不配置

注意：如果集群节点的规格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 1

\* 当前区域  温馨提示: 页面左上角切换区域。

\* 可用分区 ②

---

\* 集群名称  ②

\* 集群版本

\* Kerberos认证 ☐ 3  
未开启认证, 存在安全风险。了解更多

\* 集群类型 ☒ 分析集群 ④ ☐ 流式集群 ⑤

\* 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  [查看虚拟私有云](#)

\* 子网  [C](#)

\* 安全组  [管理安全组](#)

\* 集群高可用 ☐ ⑥

\* 集群节点

类型	实例规格 ⑦	实例数量	数据盘 ⑧	弹性伸缩 ⑨	操作
Master ⑩	4 核 16 GB   c3.xlarge.4	1	普通IO 200 GB x 1	--	--
Core ⑪	4 核 16 GB   c3.xlarge.4	1 ⑧	普通IO 100 GB x 1	--	--

配置费用: **¥3.31/小时** ⑨  
参考价格, 具体扣费请以账单为准 [了解计费详情](#)

[立即购买](#)

3.2.4 配置后点击“立即购买”进行集群的创建。等到集群创建成功后，点到集群管理界面中可以看到集群的完成状态。

## 4. 创建 MRS 流式集群并安装 Superset

### 4.1 购买 MRS 流式集群并配置 Superset

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

4.1.2 在页面中选择“购买集群”，进入“集群配置”页面。

4.1.3 在新打开的页面中填写集群的基本信息（“集群高可用”和“集群节点”配置在本次实践将采用低成本配置）：

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	mrs_demo_stream
集群版本	1.7.2
Kerberos 认证	默认值：关闭
集群类型	默认值：流式集群
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	需选择分析集群创建的安全组
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	现在配置
引导操作	现在添加，详情看 4.1.4

**注意：如果集群节点的规格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 1

\* 当前区域  温馨提示: 页面左上角切换区域。

\* 可用分区 ②

---

\* 集群名称  ② 2

\* 集群版本

\* Kerberos认证 ☐ 3  
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 ② 4

\* 组件选择

组件名	版本	描述	<input checked="" type="checkbox"/>
Kafka	0.10.2.0	分布式消息发布订阅系统	<input checked="" type="checkbox"/>
Storm	1.0.2	分布式实时计算系统	<input checked="" type="checkbox"/>
Flume	1.6.0	一个分布式、高可用、高可靠的...	<input checked="" type="checkbox"/>

5

---

\* 虚拟私有云  查看虚拟私有云 ②

\* 子网  ②

\* 安全组  管理安全组 ② 6

\* 集群高可用 ☐ ② 7

\* 集群节点

类型	实例规格 ② 8	实例数量	数据盘 ②	弹性伸缩 ②	操作
Master ②	<input type="text" value="4 核 16 GB   c3.xlarge.4"/>	<input type="text" value="1"/>	普通IO 200 GB x 1	--	--
Core ②	<input type="text" value="4 核 16 GB   c3.xlarge.4"/>	<input type="text" value="1"/> 9	普通IO 100 GB x 1	--	--

10

配置费用: ¥3.31/小时

参考价格, 具体扣费请以账单为准 了解计费详情

4.1.4 将“高级配置”设置成“现在配置”, 然后在“引导操作”内点击“添加”按钮。

\* 高级配置 ☒ 现在配置 ☐ 暂不配置

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。 查看预值

标签键  标签值

你还可以添加10个标签。

引导操作 ② ☒ 添加 你还可以添加17个引导操作。

名称	脚本路径	Master	Core	Task	参数	执行时机	失败操作
crawl	s3a://mrs-sam...	✓ Active Mast...			--	组件启动后	继续

4.1.5 引导操作的内容如下：



参数	值
名称	crawl
脚本路径	s3a://mrs-samples-bootstrap-cn-north-1/superset/superset_install.sh (华北区) s3a://mrs-samples-bootstrap-cn-east-2/superset/superset_install.sh (华东区) s3a://mrs-samples-bootstrap-cn-south-1/superset/superset_install.sh (华南区)
执行节点	Master 将 Active Master 按钮打开。
参数	默认值：(空)
执行时机	默认值：组件启动后
失败操作	继续

引导操作

\* 名称

\* 脚本路径

\* 执行节点

☒ Master
☐ Core
☐ Task

Active Master
☒

?

参数

\* 执行时机

组件启动后

\* 失败操作

继续

确定

取消

4.1.6 配置后点击“立即购买”进行集群的创建。等到集群创建成功后，点到集群管理界面中可以看到引导操作的完成状态。

## 4.2 申请弹性 IP

4.2.1 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。

4.2.2 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-stream
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

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。

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

Anti-DDoS流量清洗服务可以为华为云内资源，提供网络层和应用层的DDoS攻击防护和攻击实时告警通知。[了解更多 提升防护能力](#)

带宽名称

mrs-demo-stream

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

您还可以添加10个标签。

购买量

-

1

+

一次最多可以购买20个弹性公网IP。您还可以购买20个弹性公网IP。申请更多弹性公网IP配额请单击[申请扩大配额](#)

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

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

## 4.3 为流式集群绑定弹性 IP

4.3.1 点击 [MapReduce 服务](#) 进入集群列表页面，在“现有集群”中点击之前创建的 MRS 流式集群的名称，进入到该集群的管理页面。

您还剩余196台弹性云服务器、8个

集群列表

现有集群

历史集群

操作日志

名称	集群ID
mrs_demo_stream	443ad!

4.3.2 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_l0PEV	运行中	Core
d16d60dd-092e-4c75-93e5-6df35323e1e9_node_master1_l1XTX	运行中	Master1

4.3.3 然后在下方的菜单栏中点击“弹性 IP”菜单，然后点击“绑定弹性 IP”按钮，绑定之前创建的弹性 IP。



## 5. 使用 devcloud 编译部署程序

### 5.1 接收并构建任务

#### 5.1.1 点击

<https://devcloud.huaweicloud.com/classroom/joinhomework/4ac91fa71cd1485a8870f1c43903edb8/16c555201360412d9098a48344db266f> 进入项目接收界面，点击右侧的“接收项目”按钮，等待接收后进入到实践项目页面中。

5.1.2 然后点击左侧菜单栏中的“构建&发布->编译”页签，然后点击右边的“开始构建”按钮进行构建，可以点击项目名称来查看详细的编译构建过程。



### 5.2 添加主机授信（已添加过的可跳过该步骤）

5.2.1 点击软件开发云上面的“服务”下拉框，选中“部署”按钮进入到部署页面。然后点击“主机管理”标签，点击右侧的“添加主机”按钮进行配置。



5.2.2 在新打开的页面中填写主机的基本信息，按照以下方式填写：

参数	值
添加方式	默认值：单个添加
主机名	mrs_demo
弹性 IP	填写之前创建的流式集群弹性 IP，如：10.10.10.10
操作系统	默认值：linux
授信方式	密码授信
用户名	root
密码	创建集群时设置的密码

添加主机



添加方式：

单个添加

多个添加

\* 主机名：

mrs\_demo

\* 弹性IP：

10.10.10.10

如果没有弹性IP，请依照[弹性IP帮助文档](#)到华为云申请弹性IP

\* 操作系统：

linux

请依照[Linux配置文档](#)确认机器配置,避免授信失败

\* 授信方式：

☒ 密码授信☐ 密钥授信

\* 用户名：

root

建议修改默认的22端口或使用密钥授信方式

\* 密码：

\*\*\*\*\*

※默认操作系统是linux的主机以22端口授信，windows主机以5986端口授信。

※

※如需修改默认授信端口，或者通过SSH代理等高级功能，请点开更多。

更多



添加

取消

## 5.3 配置部署任务

5.3.1 点击左上角的“首页”，然后点击项目名称回到实践项目的页面中，点击左侧菜单栏中的“构建&发布->部署”页签，然后点击右边的“新建任务”按钮进行部署任务配置。



5.3.2 选择**非模板**任务，在新打开的页面中填写部署任务的基本信息，然后点击“确认”进行创建，下表为必填的参数列表：

参数	值
名称	weather_example
部署类型	shell 部署

选择软件包	通过界面选择添加之前构建生成的软件包：weather-crawl-mrs.jar
配置主机	通过界面选择之前配置授信的 IP，如：10.10.10.10
部署路径	/opt/
shell 脚本	<p>在脚本文件的最下方添加以下内容（<b>原内容不能删除</b>）  （<b>password 为分析集群 master 节点 root 用户密码</b>，<b>masterIP 为分析集群节点 IP</b>）</p> <pre> #初始化需要的参数值 masterIp="masterIp" analysis_password="password" zookeeper_connect_analysis="\$masterIp":2181 bootstrap_servers=`grep bootstrap.servers /opt/client/Kafka/kafka/ config/producer.properties cut -d=' ' -f2` rm -rf /opt/conf/ mkdir /opt/conf/ source /opt/client/bigdata_env #下载程序运行需要的依赖包和配置文件 wget https://obs-weather-crawler.obs-website.cn-east-2.myhwclouds.com/jsoup-1.10.3.jar mv jsoup-1.10.3.jar /opt/conf wget https://obs-devcloud.obs.cn-north-1.myhwclouds.com/weather/properties.properties #修改 flume 的配置文件，填入对应的环境参数 sed -i "s/bootstrap_servers/\$bootstrap_servers/g" /opt/properties.properties sed -i "s/zookeeper_connect_analysis/\$zookeeper_connect_analysis/g" /opt/properties.properties nohup mv properties.properties /opt/Bigdata/MRS/FusionInsight-Flume-1.6.0/flume/conf/ chown omm:wheel /opt/Bigdata/MRS/FusionInsight-Flume-1.6.0/flume/conf/properties.properties cp weather-crawl-mrs.jar /opt/Bigdata/MRS/FusionInsight-Flume-1.6.0/flume/lib/ chown omm:wheel /opt/Bigdata/MRS/FusionInsight-Flume-1.6.0/flume/lib/weather-crawl-mrs.jar chmod 755 /opt/Bigdata/MRS/FusionInsight-Flume-1.6.0/flume/lib/weather-crawl-mrs.jar #拷贝启动程序需要的分析集群的配置文件 check_user=`whoami` echo "\$check_user" </pre>

```

echo '#!/usr/bin/expect -f'>/opt/mrs.sh
echo "set password $analysis_password">>/opt/mrs.sh
echo "spawn scp -r root@$masterIp:/opt/client/Hive/config/* root@$masterIp:/opt/client/HBase/hbase/conf/* root@$masterIp:/etc/hosts /opt/conf/">>/opt/mrs.sh
echo 'expect "*password"'>>/opt/mrs.sh
echo 'send "$password\n"'>>/opt/mrs.sh
echo 'expect "*password"'>>/opt/mrs.sh
echo 'send "$password\n"'>>/opt/mrs.sh
echo 'expect "*password"'>>/opt/mrs.sh
echo 'send "$password\n"'>>/opt/mrs.sh
echo 'expect "*password"'>>/opt/mrs.sh
echo 'send "$password\n"'>>/opt/mrs.sh
echo 'interact'>>/opt/mrs.sh
chmod 755 /opt/mrs.sh
./mrs.sh
chown -R omm:wheel /opt/conf/
chmod 755 /opt/conf/*
if [ "`cat /etc/hosts |grep "$masterIp"`" == "" ];then
    cat /opt/conf/hosts |grep node >> /etc/hosts
fi
#启动爬虫程序
nohup java -cp ../weather-crawl-mrs.jar:/opt/conf:/opt/conf/jsoup-1.10.3.jar:/opt/client/Hive/Beeline/lib/*:/opt/client/HDFS/hadoop/lib/*:/opt/client/HBase/hbase/lib/*:/opt/client/Kafka/kafka/config:/opt/client/Kafka/kafka/libs/* com.huawei.mrs.crawl.WeatherCrawler > /opt/log.log 2>&1 &
sleep 10
#使用 omm 用户启动 flume 客户端
echo "exec su omm <<EOF" > /opt/startFlume.sh
echo "sh /opt/Bigdata/MRS/FusionInsight-Flume-1.6.0/flume/bin/flume-manage.sh restart" >> /opt/startFlume.sh
echo "EOF" >> /opt/startFlume.sh
chmod 755 startFlume.sh
nohup sh /opt/startFlume.sh

```

>>主机名信息可以在节点信息中查看：





- 节点信息
- 告警列表
- 标签
- 引导操作
- 运维管理

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71个节点。[申请扩大配额](#)

名称	状态	类型	IP地址
<a href="#">cf575253-fcbb-4326-9f55-5882768b9002_node_master1_XziGZ</a>	运行中	Master1	192.168.0.42
<a href="#">cf575253-fcbb-4326-9f55-5882768b9002_node_core_ZEf0g</a>	运行中	Core	192.168.0.83

>>最终效果如下图：

名称：

weather\_example

所属项目：

实战训练营之天气情况统计\_实战训练营\_基于大数据的天气情况统计\_小组op\_svc\_mrs\_container1

描述：

部署类型：

Shell部署(Linux)

容器部署

Ansible部署

运行环境：

☐ 是否安装运行环境

选择软件包：

软件包路径

/mrs\_weatherDemo\_1541579201316/20181107.1/weather-crawl-mrs.jar

选择软件包，您选择的软件包在以下的shell脚本中以{{package\_uris}}被引用

配置主机：

已选主机列表

请输入主机名或IP，按... Q

☐

No.

主机组

主机名

IP

操作

☐

1

--

mrs\_demo

10.10.10.10

批量删除

+

部署路径：

/opt

软件包部署到您的主机上的路径，您填写的内容在以下的shell脚本中以{{deploy\_path}}被引用

停止命令：

e.g. sh /usr/share/tomcat/webapps/Tomcat\_Test/stop.sh

停止命令，您填写的内容在以下的shell脚本中以{{stop\_sh}}被引用

启动命令：

e.g. sh /usr/share/tomcat/webapps/Tomcat\_Test/start.sh

启动命令，您填写的内容在以下的shell脚本中以{{start\_sh}}被引用

Shell脚本：

```
192 echo "expect \"$password\"" >>/opt/mrs.sh
193 echo "send \"$password\\n\"" >>/opt/mrs.sh
194 echo "expect \"$password\"" >>/opt/mrs.sh
195 echo "send \"$password\\n\"" >>/opt/mrs.sh
196 echo "interact" >>/opt/mrs.sh
197 chmod 755 /opt/mrs.sh
198 ./mrs.sh
199 chown -R om:wheel /opt/conf/
200 chmod 755 /opt/conf/*
201 if [ "$(cat /etc/hosts |grep "$masterIp" |wc -l)" -eq 0 ];then
202   cat /opt/conf/hosts |grep mode >> /etc/hosts
203 fi
```

高级配置

## 5.4 执行程序并查看结果

5.4.1 进入到部署详情页面，然后点击“一键部署”按钮，即可完成部署。等到部署结束后就可以在界面上看到程序的执行结果，效果如下图：

2018-11-30

华为保密信息,未经授权禁止扩散

第 17 页，共 26 页

部署日志 部署历史

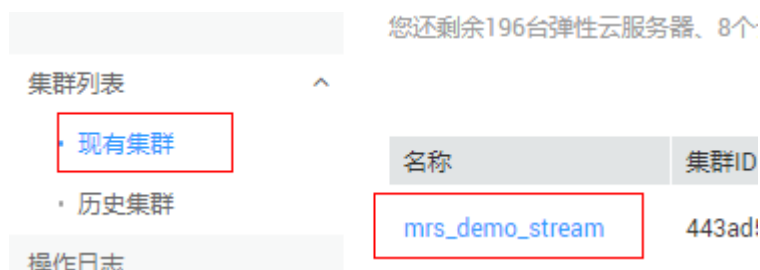
```
2 2018-10-09 16:07:57,811 [
3 "到部署路径 /opt/",
4 "传递的url是:",
5 "软件包备份到文件夹/package_bak/下",
6 "下载软件包成功",
7 "Start flume successfully, pid=31160.",
8 "root",
9 "spawn scp -r root@192.168.128.231:/opt/client/Hive/conf/* root@192.168.128.231:/opt/client/HBase/hbase/conf/* /opt/conf/",
10 "root@192.168.128.231's password:",
11 "",
12 "hiveclient.properties          0%    0    0.0KB/s  --:-- ETA",
13 "hiveclient.properties        100% 129    0.1KB/s  00:00  ",
14 "",
15 "hivemetastore-site.xml        0%    0    0.0KB/s  --:-- ETA",
16 "hivemetastore-site.xml      100% 1709    1.7KB/s  00:00  ",
17 "",
18 "hive-site.xml                 0%    0    0.0KB/s  --:-- ETA",
19 "hive-site.xml               100% 698    0.7KB/s  00:00  ",
20 "root@192.168.128.231's password:",
21 "",
22 "client.env                    0%    0    0.0KB/s  --:-- ETA",
23 "client.env                   100%  87    0.1KB/s  00:00  ",
24 "",
25 "core-site.xml                 0%    0    0.0KB/s  --:-- ETA",
26 "core-site.xml               100% 5647    5.5KB/s  00:00  ",
27 "",
28 "gc-opts.sh                    0%    0    0.0KB/s  --:-- ETA",
29 "gc-opts.sh                   100% 17KB   17.1KB/s  00:00  ",
30 "",
31 "hadoop-metrics2-hbase.properties 0%    0    0.0KB/s  --:-- ETA",
32 "hadoop-metrics2-hbase.properties 100% 1811    1.8KB/s  00:00  ",
```

5.4.2 如果想更详细的了解整个过程，可以进行代码修改然后提交到代码仓库，接着只需执行编译和部署的执行按钮就可重新查看结果。也可以通过 putty 等工具连接到 MRS 集群上，通过 MRS 客户端进行其他操作。具体操作手册参考帮助文档：[https://support.huaweicloud.com/usermanual-mrs/mrs\\_01\\_0081.html](https://support.huaweicloud.com/usermanual-mrs/mrs_01_0081.html)

## 6. 使用 superset 展示数据

### 6.1 为流式集群添加安全规则

6.1.1 进入 MRS 集群列表，点击之前创建的流式集群名称，进入到集群管理页面。



6.1.2 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_1cPEV	运行中	Core
d16d60dd-092e-4c75-93e5-6df35323e1e9_node_master1_IIXTX	运行中	Master1

6.1.3 点击“安全组”，再点击 ID 进入到安全组设置界面。

云服务器控制台

总览 弹性云服务器

专属主机 云服务器备份 裸金属服务器 云硬盘 专属分布式存储

云硬盘 网卡 **安全组** 弹性IP 监控 标签

更改安全组

^ mrs\_mrs\_zzr0\_Adty

出方向规则 1 入方向规则 11 ID: db0e61ab-fb03-4e24-8396-1d5b9a24f3c9

方向	类型	协议
入方向	IPv4	Any
入方向	IPv4	Any

6.1.4 选择入方向，点击“添加规则”，在端口范围栏里填上“38088”，点击确定。

添加入方向规则

安全组 mrs\_mrs\_demo\_stream\_WpoD

协议/应用	端口和源地址	描述	操作
TCP	端口 38088 源地址 IP地址 0.0.0.0 / 0		复制 删除

+ 增加1条规则 您还可以增加9条规则

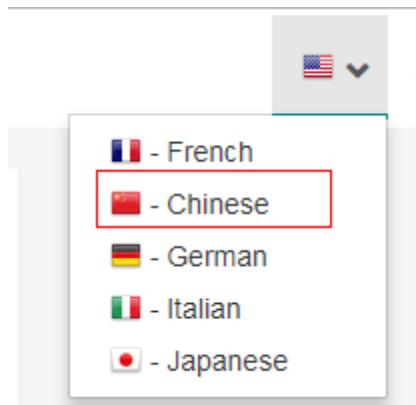
确定 取消

## 6.2 使用 superset

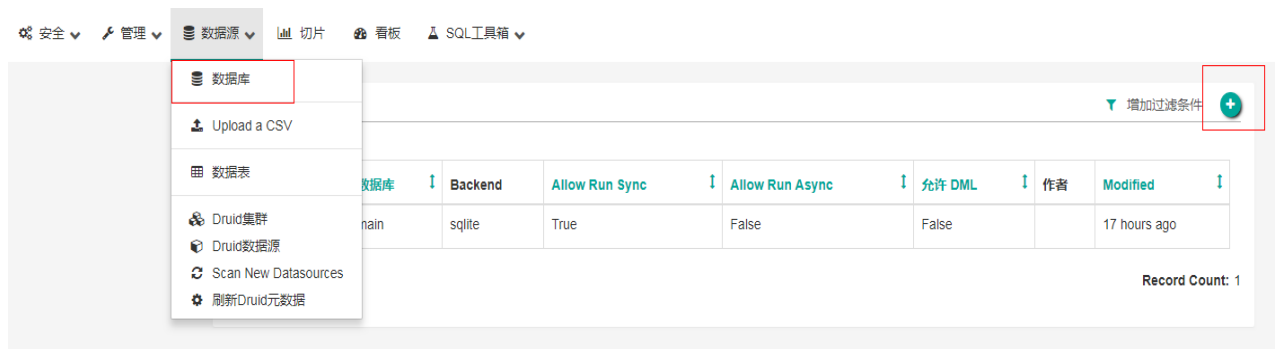
6.2.1 当爬虫程序启动后，可在浏览器输入如下地址登录到 superset 工具中。

http://弹性 IP:38088

6.2.2 账号为：admin，密码为：Admin12!。登录后可在右上角将语言改成中文。



6.2.3 在顶部菜单栏点击“数据源->数据库”，进入到数据页面，点击右上角的“添加新纪录”按钮，进行 Hive 的关联。



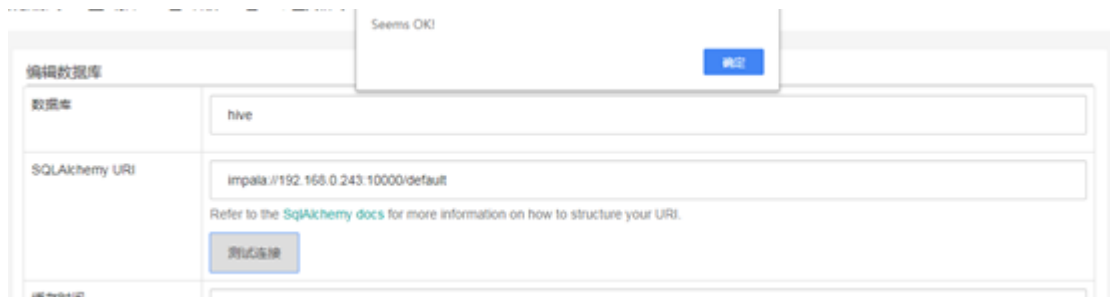
6.2.4 需要修改的配置参数参考下面表格，其余参数保持默认值（**masterIP** 为分析集群 **master** 节点 **ip**）。

参数	值
数据库	hive
SQLAlchemy URI	impala:// <b>masterIP(192 开头的 IP):10000/default</b>
扩展	<pre>{   "metadata_params": {},   "engine_params": {<b>"connect_args":{"auth_mechanism":</b> <b>PLAIN","kerberos_service_name":"hive"}}</b> }</pre>
在 SQL 工具箱中公开	勾选

## 添加数据库

数据库	hive
SQLAlchemy URI	impala://192.168.0.243:10000/default Refer to the <a href="#">SQLAlchemy docs</a> for more information on how to structure your URI. 测试连接
缓存时间	缓存时间
扩展	<pre>{   "metadata_params": {},   "engine_params": {"connect_args":{"auth_mechanism":"PLAIN","kerberos_service_name":"hive"}} }</pre> JSON string containing extra configuration elements. The <code>engine_params</code> object gets unpacked into the <code>sqlalchemy.create_engine</code> call, while the <code>metadata_params</code> gets unpacked into the <code>sqlalchemy.MetaData</code> call.
在SQL工具箱中公开	<input checked="" type="checkbox"/> 在SQL工具箱中公开这个数据库
Allow Run Sync	<input checked="" type="checkbox"/> 允许用户运行同步查询，这是默认值，可以很好地处理在web请求范围内执行的查询(<~1分钟)

配置完成后点击“测试连接”按钮，当显示“Seems OK!”代表连接成功，然后点击“保存”。



6.2.5 点击上方的“SQL 工具箱->SQL 编辑器”，然后在左侧选择相应数据库和表。选择后有报错可忽略。

SQL 编辑器

Database: hive  
Schema: default  
Add a table (1)

hive\_weather

- hive\_weather.rowkey
- hive\_weather.id
- hive\_weather.province
- hive\_weather.city
- hive\_weather.zone
- hive\_weather.time
- hive\_weather.maxtemperature
- hive\_weather.mintemperature
- hive\_weather.weather
- hive\_weather.windpower

```
select hive_weather.province, hive_weather.city as city, hive_weather.zone as zone, hive_weather.time as time, hive_weather.maxtemperature as maxtemperature,
hive_weather.mintemperature as mintemperature, hive_weather.weather as weather, hive_weather.windpower as windpower from hive_weather
```

Run Query Save Query

Results Query History Preview for hive\_weather

province	city	zone	time	maxtemperature	mintemperature	weather	windpower
北京	北京	北京	2011-01-01	0	-9	晴 /晴	无持续风向 <3级 /无持续风向 <3级
北京	北京	北京	2011-01-02	-2	-7	多云 /晴	无持续风向 <3级 /无持续风向 <3级
北京	北京	北京	2011-01-03	1	-8	晴 /晴	北风 3-4级 /无持续风向 <3级
北京	北京	北京	2011-01-04	-1	-11	晴 /晴	无持续风向 <3级 /无持续风向 <3级
北京	北京	北京	2011-01-05	-1	-8	晴 /晴	北风 4-5级 /北风 3-4级
北京	北京	北京	2011-01-06	0	-10	晴 /晴	无持续风向 <3级 /无持续风向 <3级
北京	北京	北京	2011-01-07	1	-7	晴 /多云	无持续风向 <3级 /无持续风向 <3级
北京	北京	北京	2011-01-08	1	-8	多云 /晴	北风 4-5级 /北风 4-5级

6.2.6 接着输入如下的 SQL 语句进行查询：

```
select hive_weather.province as province,hive_weather.city as city,hive_weather.zone as
zone,hive_weather.time as time,hive_weather.maxtemperature as maxtemperature,
hive_weather.mintemperature as mintemperature,hive_weather.weather as
weather,hive_weather.windpower as windpower from hive_weather
```

```
1 select hive_weather.province as province,hive_weather.city as city,hive_weather.zone as zone,hive_weather.time as time,hive_weather.maxtemperature as maxtemperature,
2 hive_weather.mintemperature as mintemperature,hive_weather.weather as weather,hive_weather.windpower as windpower from hive_weather
3
```

Run Query

Save Query

Results

Query History

Preview for hive\_weather

Visualize

CSV

province	city	zone	time	maxtemperature	mintemperature	weather	windpower
北京	北京	北京	2011-01-01	0	-9	晴/晴	无持续风向 ≤3级/无持续风向 ≤3级
北京	北京	北京	2011-01-02	-2	-7	多云/阴	无持续风向 ≤3级/无持续风向 ≤3级
北京	北京	北京	2011-01-03	1	-8	晴/晴	北风 3-4级/无持续风向 ≤3级
北京	北京	北京	2011-01-04	-1	-11	晴/晴	无持续风向 ≤3级/无持续风向 ≤3级
北京	北京	北京	2011-01-05	-1	-8	晴/晴	北风 4-5级/北风 3-4级
北京	北京	北京	2011-01-06	0	-10	晴/晴	无持续风向 ≤3级/无持续风向 ≤3级

6.2.7 然后点击“Visualize”按钮按照下图方式进行配置。然后点击“Visualize”按钮，等待两分钟，会出现最终的柱状图结果，此时显示的结果是整个北京的总温度。

## Visualize



Chart Type

分布-柱状图



Datasource Name

mrs

column	is_dimension	is_date	agg_func
province	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Select... ▼
city	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Select... ▼
zone	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Select... ▼
time	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Select... ▼
maxtemperature	<input checked="" type="checkbox"/>	<input type="checkbox"/>	SUM(x) x ▼
mintemperature	<input checked="" type="checkbox"/>	<input type="checkbox"/>	SUM(x) x ▼
weather	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Select... ▼
windpower	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Select... ▼

Visualize

6.2.8 接下来就可以根据自己的要求进行数据调整和展示。比如做一个北京地区的历史最高温度曲线图，可以参考如下配置：

参数	值
图表类型	时间序列-折线图
Time-since	7 years ago
Group by	清空
筛选	<ol style="list-style-type: none"> <li>1. 选择“zone”</li> <li>2. 选择“==”</li> <li>3. 填入“北京”</li> </ol>

数据源

mrs

图表类型

时间序列-折线图

Time

时间字段

time

Time Grain

Time Column

Since

7 years ago

Until

now

Query

Metrics

sum\_\_maxtemperature

Group by

Select 8

Series limit

50

Sort By

Select 2

☒ Sort Descending

Chart Options

X Axis

Y Axis

Advanced Analytics

Annotations

Select a annotation layer

SQL

筛选

zone

==

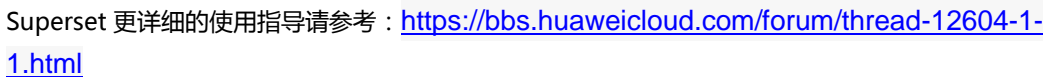
北京

6.2.9 点击上方的“Run Query”按钮，如果当前有在跑的任务，先点击“stop”按钮，再点击“Run Query”按钮。





### 6.2.10 最终结果如图：



完成步骤6.2.10，将最终结果截图并在打卡程序附上superset网站的地址。待第二天验收完成前不要清理集群和弹性IP资源。清理资源会统一发送消息通知。

## 8. 附录

未释放的集群资源也可用于附加题的操作部署或者练习前一期大数据实践训练营的内容。

第一期大数据实践训练营包含了HDFS、HBase、Hive、Spark、Mapreduce、kafka、flume、storm等主流大数据组件的练习，不仅有客户端使用练习，还有二次开发的练习，前一期活动地址：

<https://bbs.huaweicloud.com/forum/forum.php?mod=viewthread&tid=8940&page=1#pid27501>

操作手册地址：<https://obs-devcloud.obs-website.cn-north-1.myhwclouds.com/>