

文档版本号：V1.1.0-20200407

**佳都科技**

**机器学习计算引擎部署文档**

**V1.0**

**佳都新太科技股份有限公司**

**二零二零年十二月**

**目录**

[保密申明 3](#_Toc21343)

[1 前言 4](#_Toc30192)

[1.1 概述 4](#_Toc23528)

[1.2 适用范围 4](#_Toc24789)

[1.3 部署环境 4](#_Toc7423)

[1.4 相关术语 4](#_Toc31070)

[2 部署软件包 5](#_Toc7319)

[2.1 主部署包说明 5](#_Toc2331)

[3 部署环境准备 6](#_Toc885)

[3.1 部署服务器 6](#_Toc19811)

[3.2 系统参数调整 6](#_Toc6803)

[3.2.1 性能参数调整【可选】 7](#_Toc28742)

[3.2.1.1 内存设置参数调整 7](#_Toc15573)

[3.2.1.2 关闭文件更新时间、关闭swap文件交换 7](#_Toc7648)

[4 部署步骤 9](#_Toc22409)

[4.1 部署必看 9](#_Toc16736)

[4.2 上传软件到服务器 9](#_Toc30674)

[4.3 机器学习计算引擎部署 9](#_Toc3915)

[4.3.1 检查环境 9](#_Toc5220)

[4.3.2 安装模块 11](#_Toc28347)

[4.3.3 验证部署 11](#_Toc32325)

[5 备注 13](#_Toc8615)

文档控制页

|  |  |  |  |
| --- | --- | --- | --- |
| 内容 | 版本 | 时间 | 编写人 |
| 创建 | V1.4 | 2020年12月14日 | 张燕 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 保密申明

本文档内资料，用于佳都科技内部培训、施工、交流参考指引。未经同意。任何人不得以任何方式向第三方组织、机构、人员直接或间接拷贝、公布信息内容，亦不得将文档信息用于商业目的以外的其他用途，佳都科技保留一切权力。

# 

# 前言

## 概述

本文档针对多维数据融合-机器学习计算引擎的部署进行介绍。

## 适用范围

需要部署多维数据融合-机器学习计算引擎的项目。

## 部署环境

本文档以服务器安装Centos7.2 64位操作系统、JDK 1.8、Python2.7作为基础，进行软件模块的部署进行介绍。

## 相关术语

|  |  |
| --- | --- |
| **术语** | **解释** |
| JDK | [Java](http://baike.baidu.com/subview/29/12654100.htm" \t "_blank)语言的开发工具包，包含了JAVA的运行环境，工具和基础的类库 |
| ElasticSearch | ElasticSearch是一个开源的近实时全文搜索引擎 |
| Python | Python是一种跨平台的计算机程序设计语言,是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言 |

# 部署软件包

## 主部署包说明

主要软件包有：data-algorithm-1.0.0-release.tar.gz

压缩包根目录下包含如下目录：

|  |  |
| --- | --- |
| **目录** | **描述** |
| sh | 部署时使用的脚本 |
| conf | common.conf : 基础配置，本系统的基础配置会自动从nacos获取  log4j-driver.properties,log4j-executor.properties:log配置  model.sql：算法相关sql配置  relation.conf：关系代码配置 |
| config-template | 子系统使用的配置文件模板 |
| lib | 子系统的相关jar包 |
| bin | 子系统的相关运行脚本 |
| soft | 子系统部署时依赖的第三方软件 |

# 部署环境准备

## 部署服务器

在项目提供的硬件上，安装**Centos7.2**以上的操作系统，并安装**JDK 1.8以上**的JAVA运行环境, Centos 7自带P**ython2.7**运行环境。集群对应的**spark 2.4**客户端,集群对应的**hive 3.1**客户端

**注意：**

验证JAVA环境：

java -version

命令返回的jkd版本要在1.8.0\_60以上，如java version "1.8.0\_60"

验证Python环境：

python --version

命令返回的python版本要在2.7.0以上，如Python 2.7.5

验证SPARK 客户端环境：

spark-submit --version

命令返回的spark版本如2.4.0

验证hive客户端环境：

hive --version

命令返回的hive版本要求在3.0.0以上，如Hive 3.1.0.3.1.5.0-152

**如果版本不对，请按需要安装对应的软件版本**

## 系统参数调整

**如下参数必要调整**

调整方式按照《视频云系统部署重要参数调整.doc》文档中2.1必须调整的调整章节

* 设置可打开的最大文件个数、放开最大可锁定内存限制
* 安装并配置时钟同步
* **配置主机名和IP映射**

**项目要求nacos作为配置中心，需要配置在/etc/hosts中增加如下内容**

**注意：需要将xxx.xxx.xxx.xxx替换为实际的安装nacos的服务器IP**

**xxx.xxx.xxx.xxx nacos-center.v-base**

* 关闭防火墙
* 修改selinux设置
  + 1. **性能参数调整【可选】**

#### 修改文件打开最大数

ulimit -n 65535

**ps: spark集群各服务器以及部署引擎的服务器都需要修改**

#### 内存设置参数调整

注意：此步骤为可选步骤，属于性能调优部分，生产环境建议都配上，测试环境可选。

vi /etc/sysctl.conf

增加：

vm.swappiness=0 # 设置尽可能利用物理内存，在用尽物理内存情况下才使用虚拟内存

vm.max\_map\_count=655360 # 放开map counts的限制

执行如下命令使配置马上起效：

sysctl -w vm.max\_map\_count=655360

#### 关闭文件更新时间、关闭swap文件交换

注意：此步骤为可选步骤，属于性能调优部分，生产环境建议都配上，测试环境可选。

vi /etc/fstab

1. 注释包含swap的行

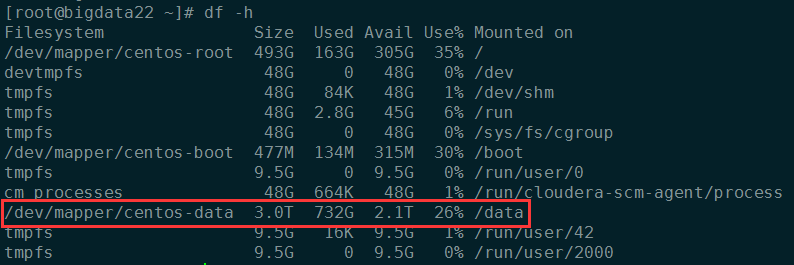
# /dev/mapper/centos-swap swap swap defaults 0 0

1. 关闭文件更新时间，若有多个数据盘需相应配置

/dev/mapper/centos-data / xfs noatime,nodiratime 0 0

一般来说，只需要在数据目录配置noatime,nodiratime，若此前相关目录已配置为default，需要把defaults修改为noatime,nodiratime，可通过以下命令查看数据目录的设备名称：

df -h



**注意：**必须重启服务器，才能使此配置起效，可在部署完毕后再重启。

**异常情况处理：**

修改完此文件重启服务器后，发现根目录下的所有文件无法修改，如用“touch xxx”命令创建文件失败并报为只读文件系统。查看/etc/fatab文件内容，若根目录的配置被注释了，说明是此错误配置造成的，需要执行如下命令使根目录恢复为可读写状态：

mount -o remount, rw /

然后修改/etc/fstab文件，把根目录的配置的注释去掉。

# 部署步骤

## 部署必看

* **机器学习计算引擎以spark集群作为底层运算框架，安装前请确认，有spark，hadoop，hive集群环境(此环境由第三方厂家提供)**
* **安装机器学习计算引擎的服务器需要安装spark和hive客户端，用于提交任务到集群计算，此客户端由提供集群的第三方厂家提供和安装**。
* **安装机器学习计算引擎的服务器需要安装任务调度的（job-engine）模块，用于调起算法任务；如果在安装机器学习计算引擎之前就安装好了任务调度的（job-engine）模块或者机器学习计算引擎的安裝路径有修改（即安装环境变量，默认为ALG\_DEPLOY\_HOME），则需要将任务调度的（job-engine）模块重启，以获取机器学习计算引擎的环境变量，否则 机器学习计算引擎的任务则会调度失败。**
* **4.3.3 上传设备关系数据到hive，非常重要，如果没有按照说明正确上传设备关系到hive，引擎中的很多模型，将无法正确执行。v1.0.0版本没有提供界面做上传设备关系的功能，只能通过4.3.3的方式人工上传。如果对此地方有质疑，请联系产品经理，此处对齐功能时，已和产品确认，不提供界面，通过指引文档的方式，人工操作。**

## 上传软件到服务器

上传data-algorithm-1.0.0-release.tar.gz到服务器，如/tmp目录，解压到安装目录，如/opt：

su - root

tar -zxf /tmp/data-algorithm-1.0.0-release.tar.gz -C **/opt**

**注意：**

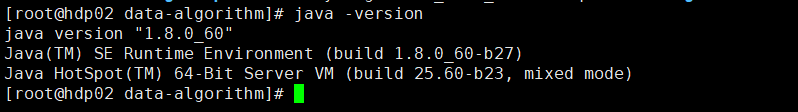
本文档默认/opt目录下此前未安装过data-algorithm，若此前已安装过data-algorithm，需要如下操作：

mv /opt/data-algorithm /opt/data-algorithm-bak，备份之前按照的目录，然后重新安装

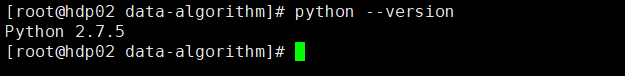
## 机器学习计算引擎部署

### 检查环境

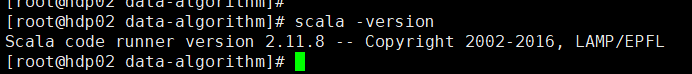
1. jdk已安装1.8版本以上



1. python 安装2.7版本以上

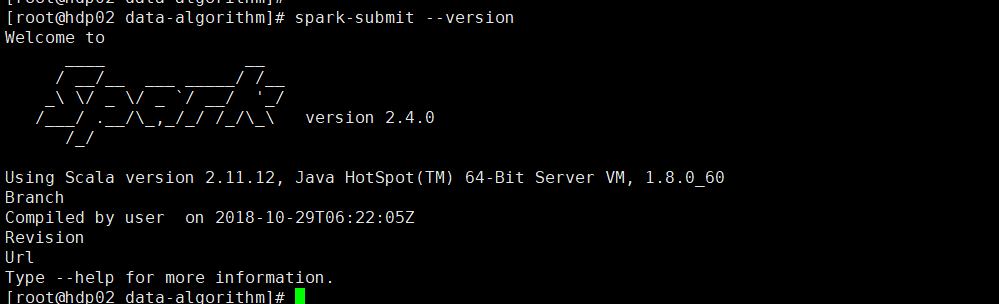


1. scala 安装2.11.8版本



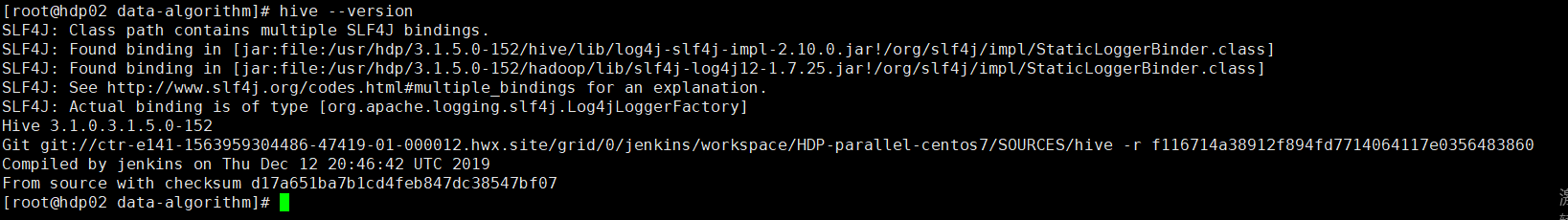
**ps: 如果没安装，模块安装时会检查并自动安装，建议提前检查版本并安装**

1. spark 客户端安装2.4.0版本



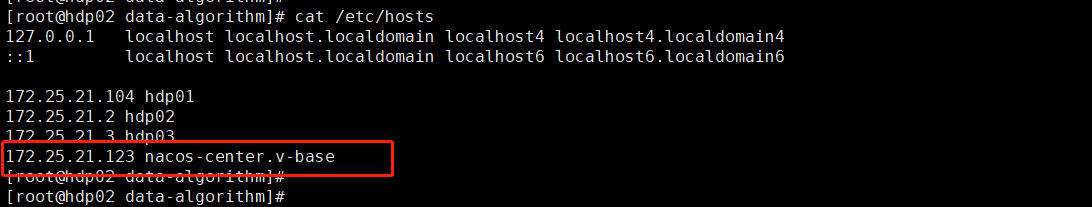
**ps: 如果没安装，请联系提供spark集群的厂家，安装客户端**

1. hive 客户端安装3.1.0版本



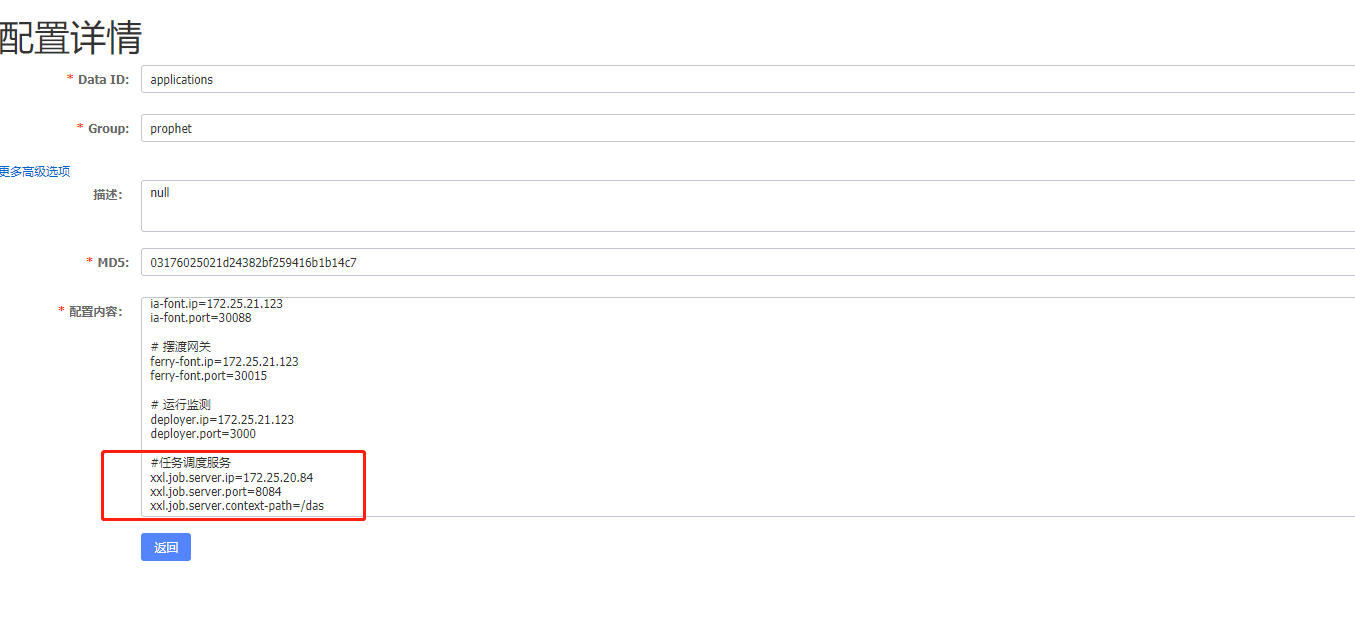
**ps: 如果没安装，请联系提供spark集群的厂家，安装客户端**

1. 检查nacos配置
   1. nacos 域名是否配置在hosts中



* 1. 检查nacos配置中心依赖配置是否已配置

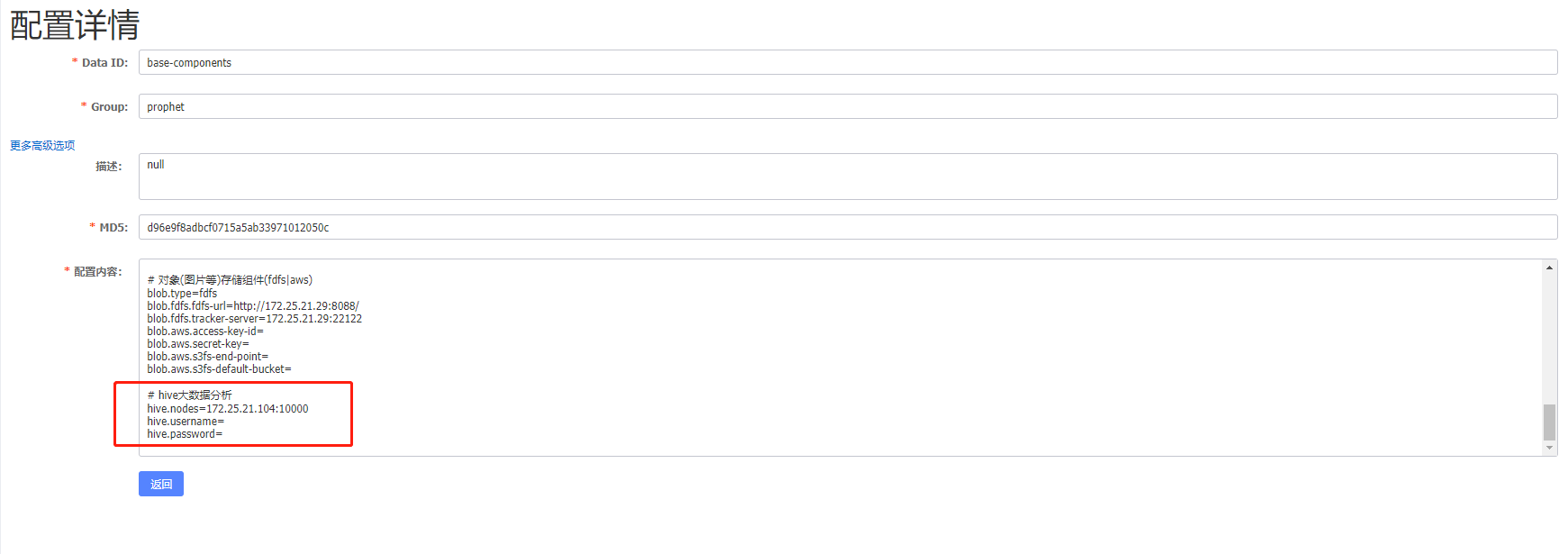
1.xxl-job(job-console)是否已经配置



1. 检查ES地址是否已经配置

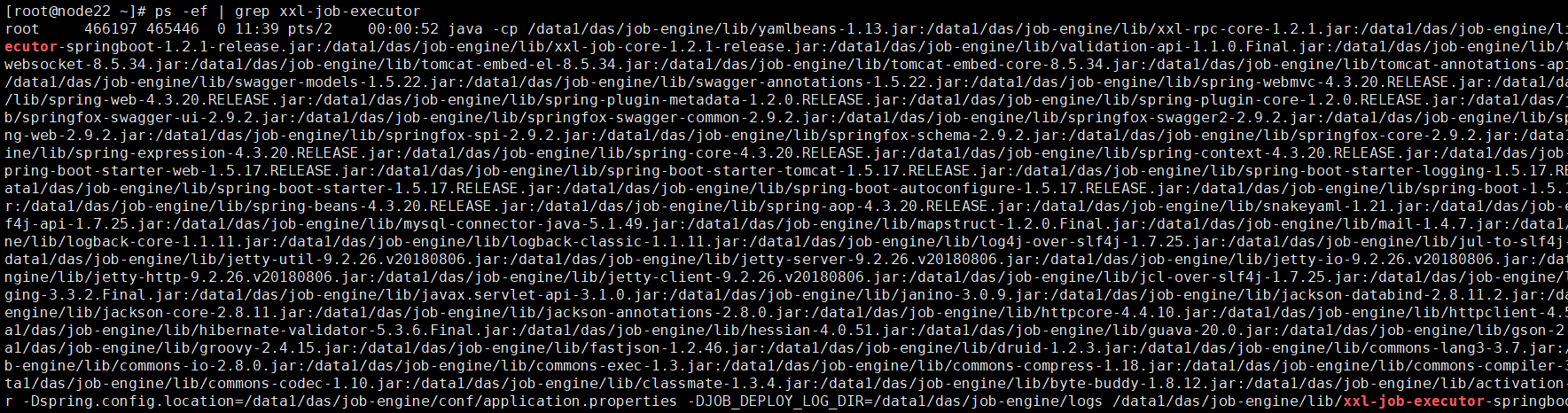


1. 检查HIVE的hive server 的地址是否已经配置（账号密码可以不配置）



HIVE server的地址可以咨询对应厂家获得！

1. 任务调度xxl-engine模块已安装



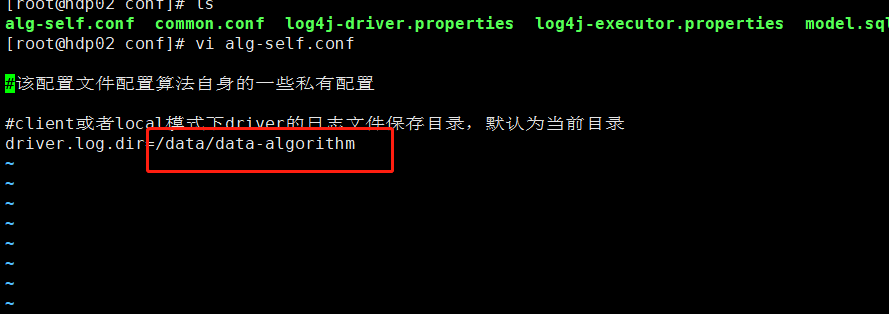
### 安装模块

1. 修改配置文件

cd conf

vi alg-self.conf

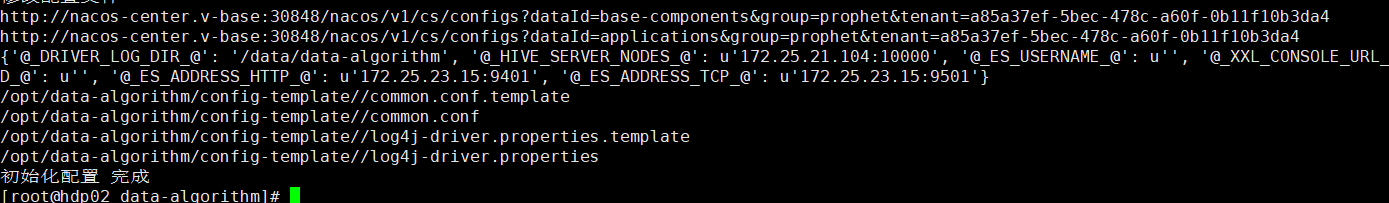
修改日志配置目录



1. 执行安装脚本

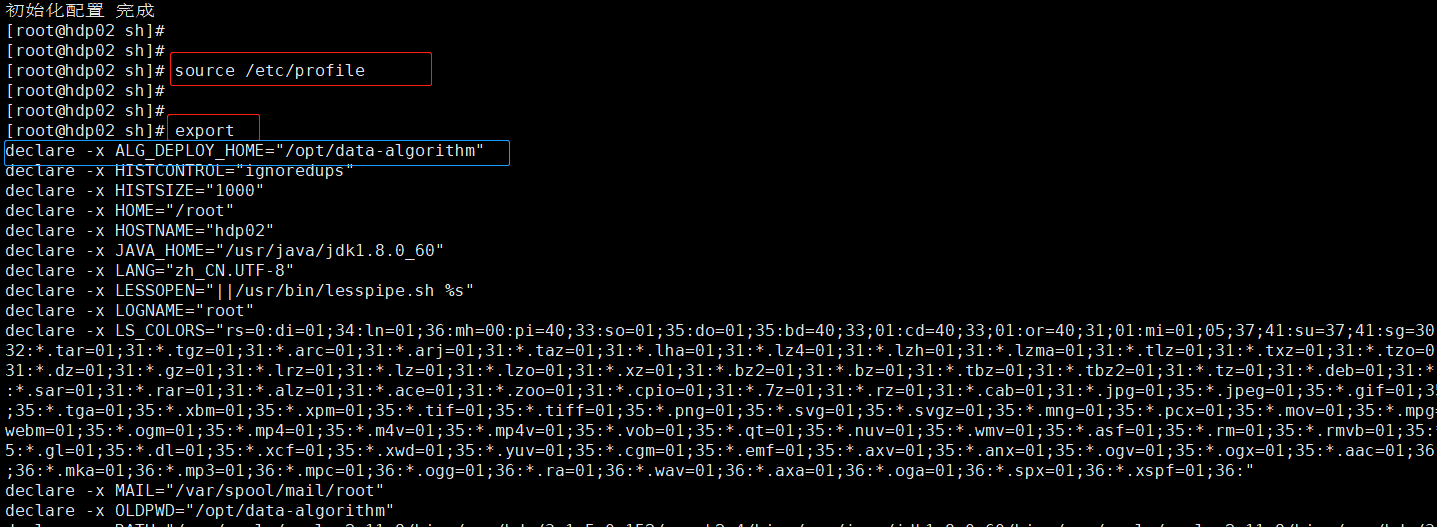
cd /opt/data-algorithm/sh

sh install.sh



1. 更新环境变量

source /etc/profile



### 上传设备关系数据到hive

1. 生成设备到设备的关系列表文件(设备与设备之间的关系，请各上线的地市找客户自行获取)。

格式为:[设备1],[设备1的类型],[设备2],[设备2的类型]

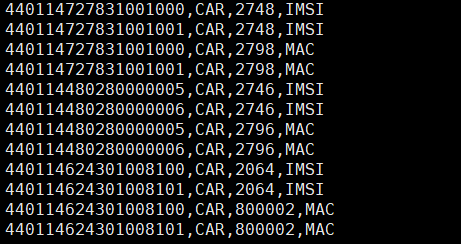
**这里有个特别注意的地方，设备1和设备2由优先关系，比如采集车牌的设备和采集imsi的设备，采集车牌设备为设备1，采集imsi的设备为设备2。**

**设备优先级关系如下：**

|  |  |
| --- | --- |
| **设备1** | **设备2** |
| **采集手机号的设备(wifi 设备)** | **采集IMEI的设备(wifi 设备)** |
| **采集手机号的设备(wifi 设备)** | **采集IMSI的设备(wifi 设备)** |
| **采集手机号的设备(wifi 设备)** | **采集MAC的设备(wifi 设备)** |
| **采集IMEI的设备(wifi 设备)** | **采集IMSI的设备(wifi 设备)** |
| **采集IMEI的设备(wifi 设备)** | **采集MAC的设备(wifi 设备)** |
| **采集IMSI的设备(wifi 设备)** | **采集MAC的设备(wifi 设备)** |
| **采集车牌的设备(卡口设备)** | **采集IMSI的设备(wifi 设备)** |
| **采集车牌的设备(卡口设备)** | **采集MAC的设备(wifi 设备)** |
| **采集车牌的设备(卡口设备)** | **采集IMEI的设备(wifi 设备)** |
| **采集车牌的设备(卡口设备)** | **采集手机号的设备(wifi 设备)** |
| **采集人脸的设备** | **采集IMSI的设备(wifi 设备)** |
| **采集人脸的设备** | **采集IMEI的设备(wifi 设备)** |
| **采集人脸的设备** | **采集MAC的设备(wifi 设备)** |
| **采集人脸的设备** | **采集手机号的设备(wifi 设备)** |

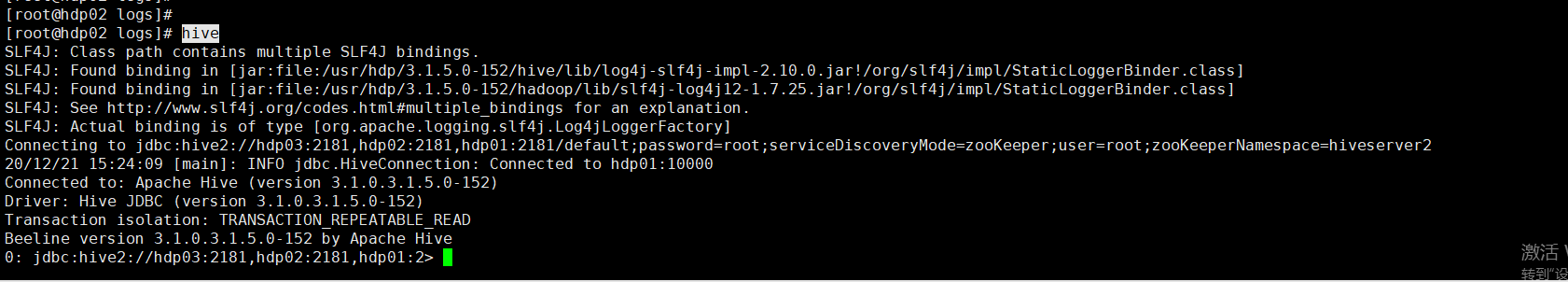
|  |  |
| --- | --- |
| **设备类型** | **设备值** |
| **采集手机号的设备(wifi 设备)** | **PHONE** |
| **采集IMEI的设备(wifi 设备)** | **IMEI** |
| **采集IMSI的设备(wifi 设备)** | **IMSI** |
| **采集MAC的设备(wifi 设备)** | **MAC** |
| **采集人脸的设备** | **PERSON** |
| **采集车牌的设备(卡口设备)** | **CAR** |

**如下示例（请注意大小写）：**



1. 将上述数据，存储到device.txt中，并将该文件上传到spark集群的某一个节点上。
2. 将device.txt上传的hdfs中
3. 连接hive客户端

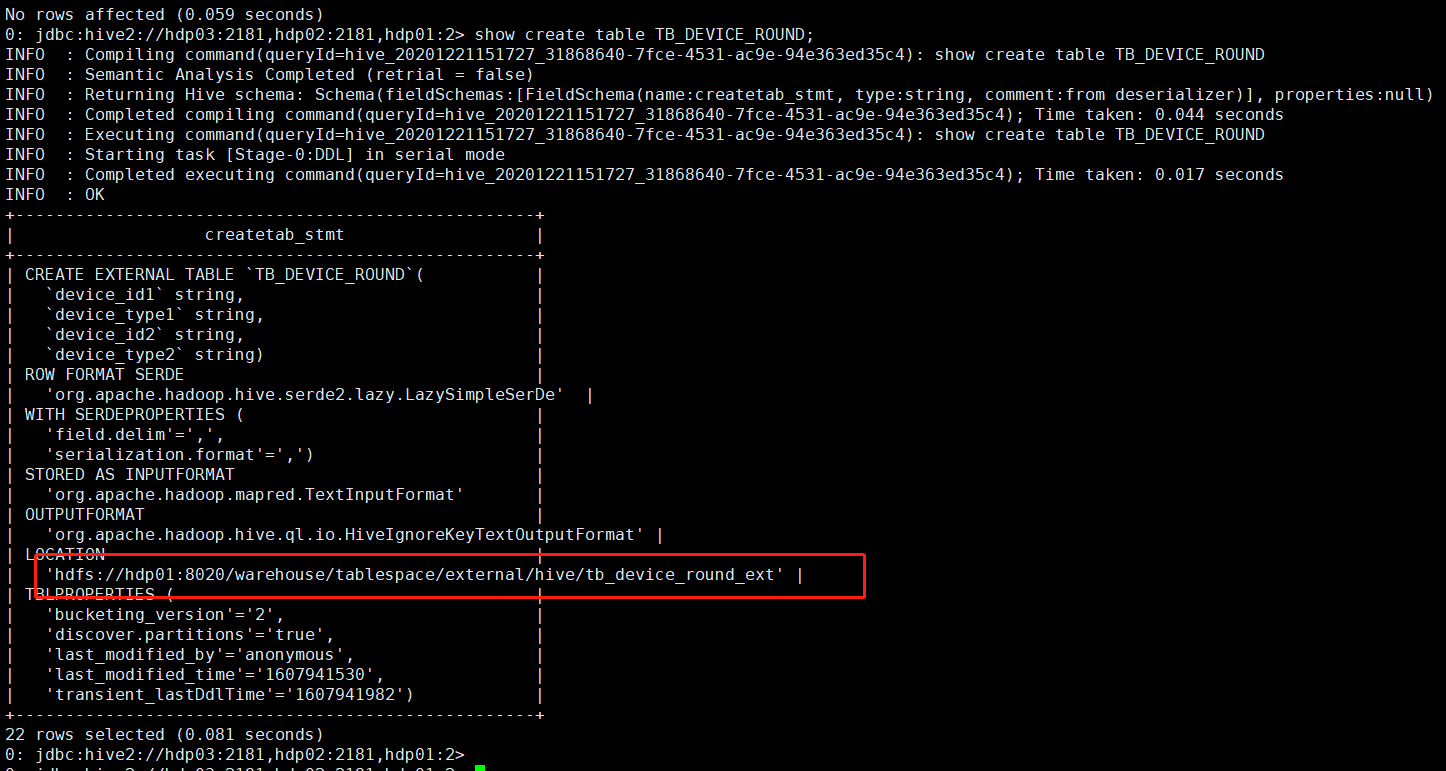
hive



1. 找到设备表对应的hdfs文件路径，

show create table TB\_DEVICE\_ROUND;

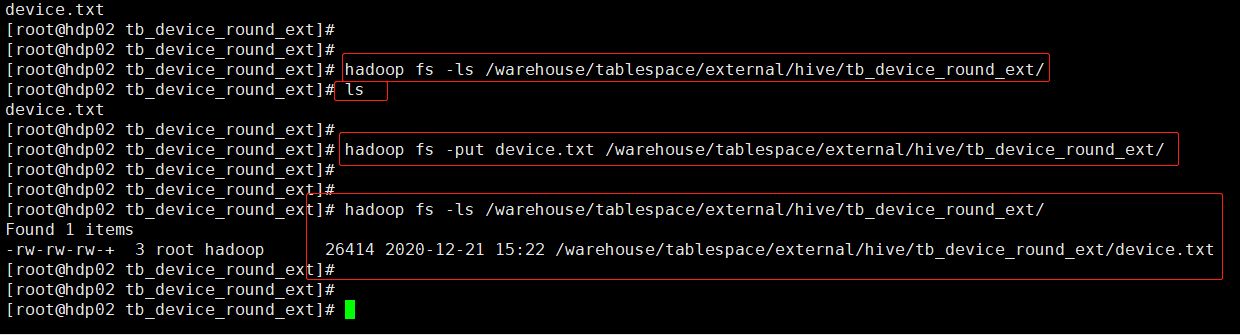
如下图：



1. 上传device.txt 到对应的hdfs 路径下

hadoop fs -put device.txt [hdfs 路径]

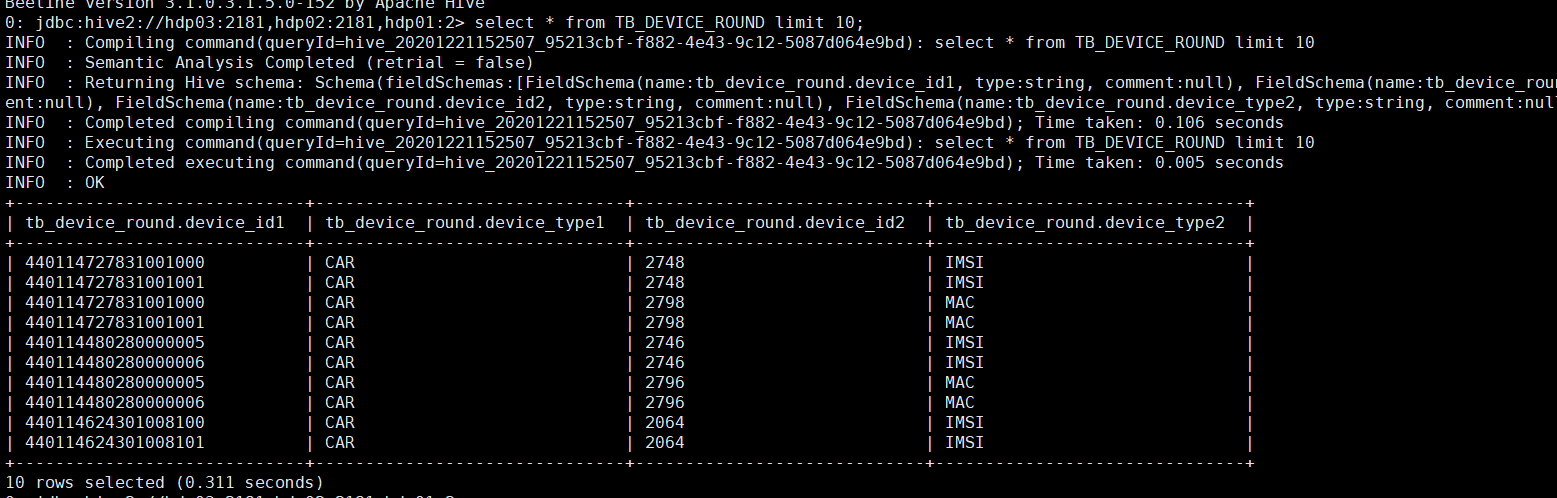
操作举例，如下图：



1. 验证是否成功

连接hive客户端

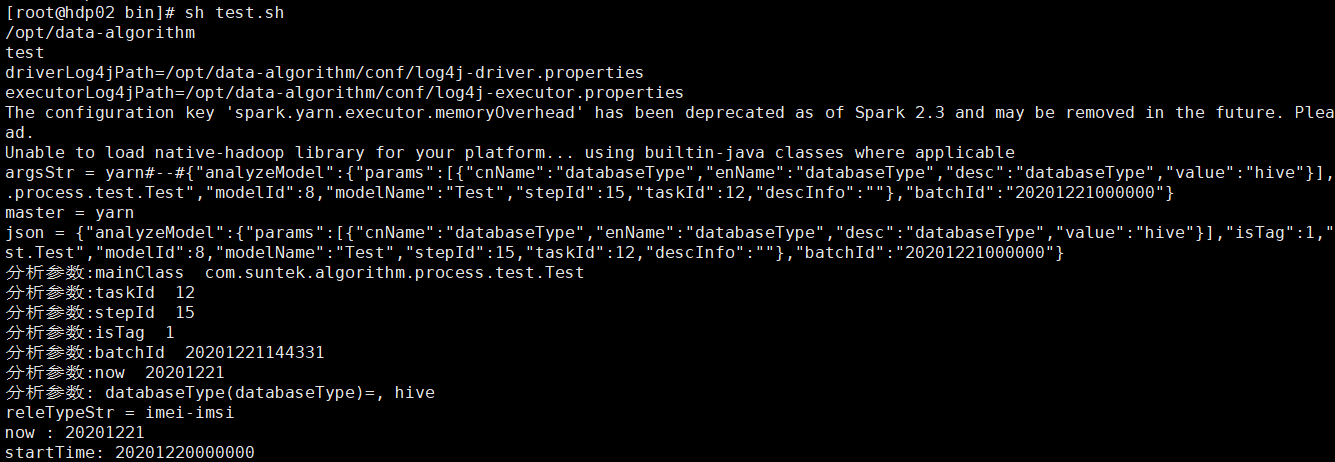
执行select \* from TB\_DEVICE\_ROUND limit 10;可以查到数据，如下图，则证明导入成功。



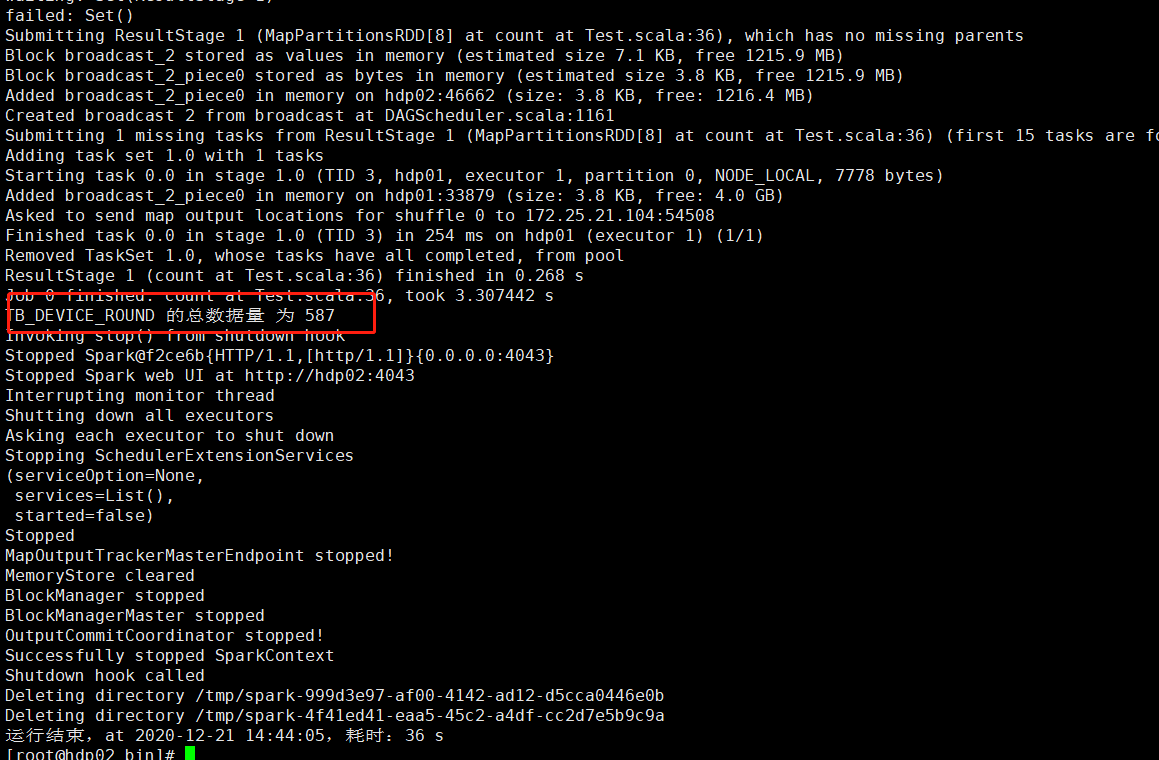
### 验证部署

cd /opt/data-algorithm/bin

执行sh test.sh，如下图



查看结果如下图，没有报错。则说明引擎部署成功。



# 备注