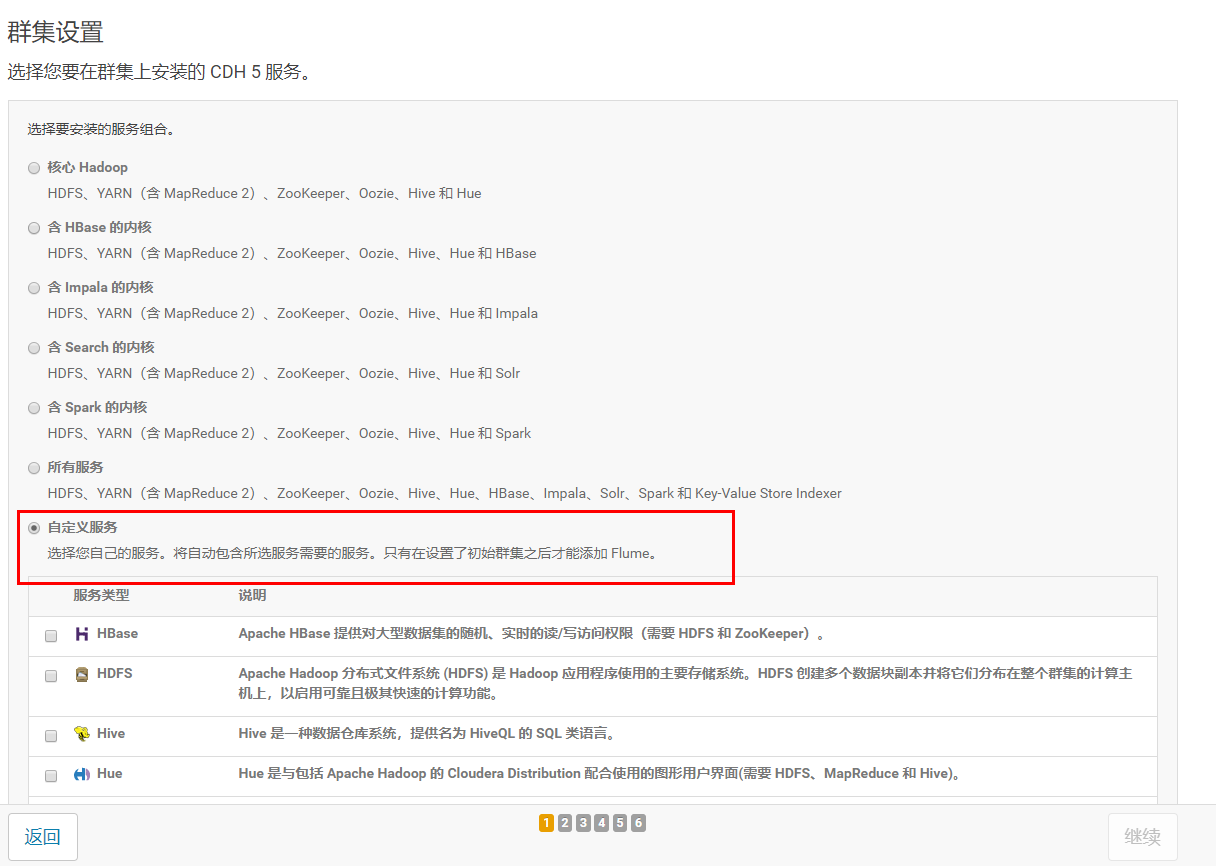
数仓之数据采集模块

## 2.1 HDFS、YARN、Zookeeper安装

### 2.1.1 选择自定义安装



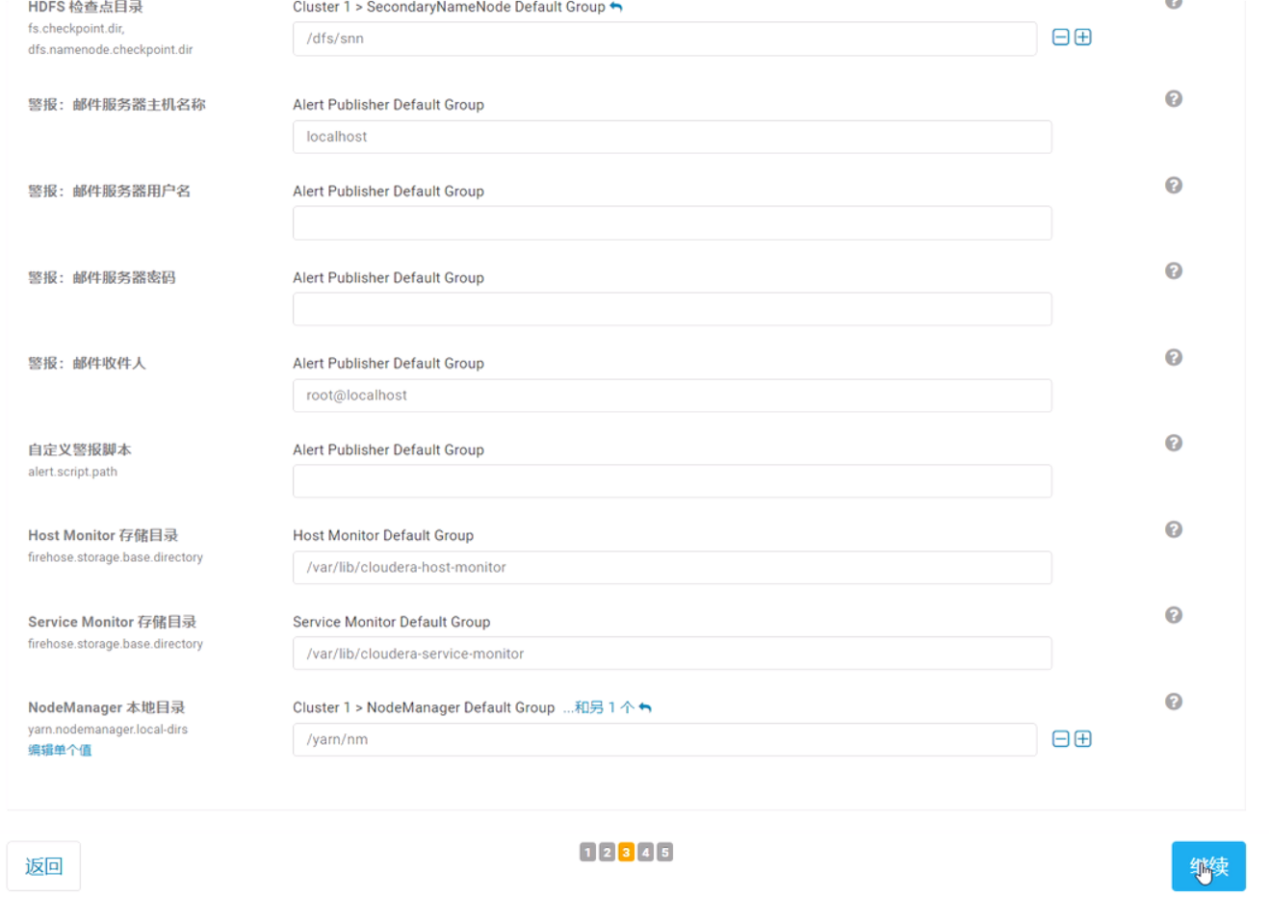
### 2.1.2 选择安装服务



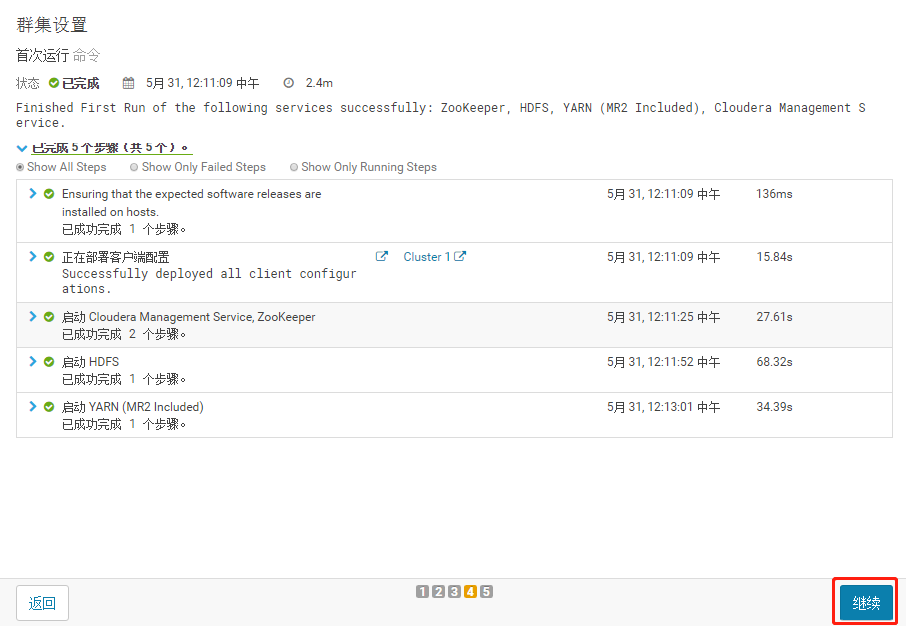
### 2.1.3 分配节点



### 2.1.4 集群设置全部选默认即可

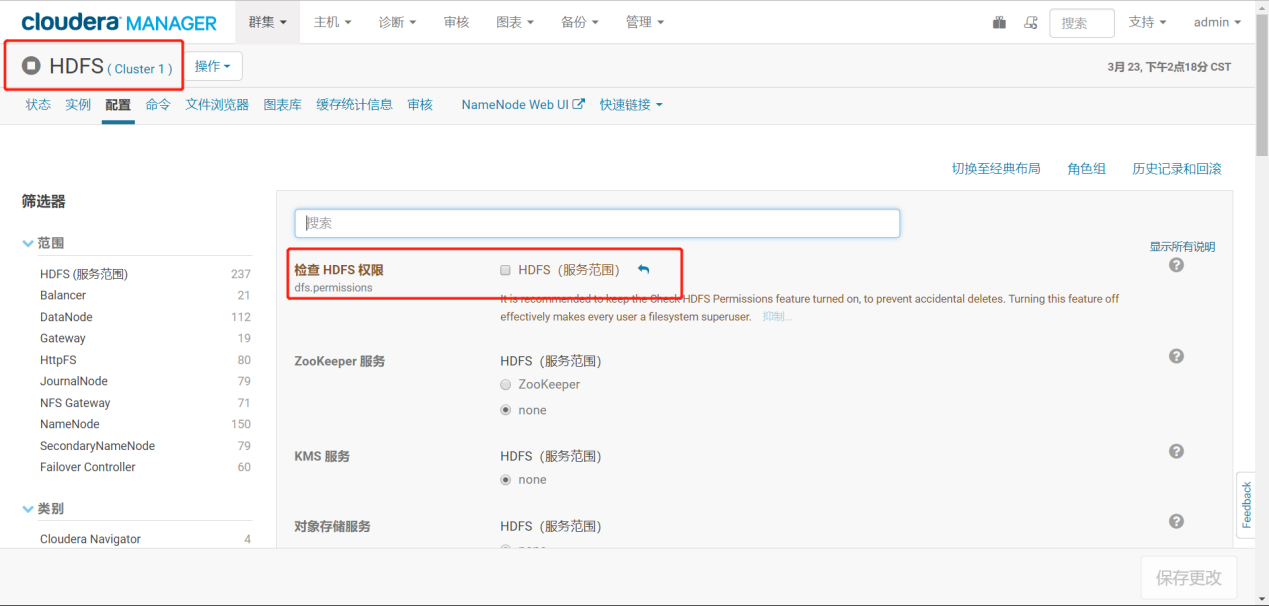


### 2.1.5 自动启动进程



### 2.1.6 修改HDFS的权限检查配置

关闭HDFS中的权限检查：dfs.permissions。



### 2.1.7 配置Hadoop支持LZO

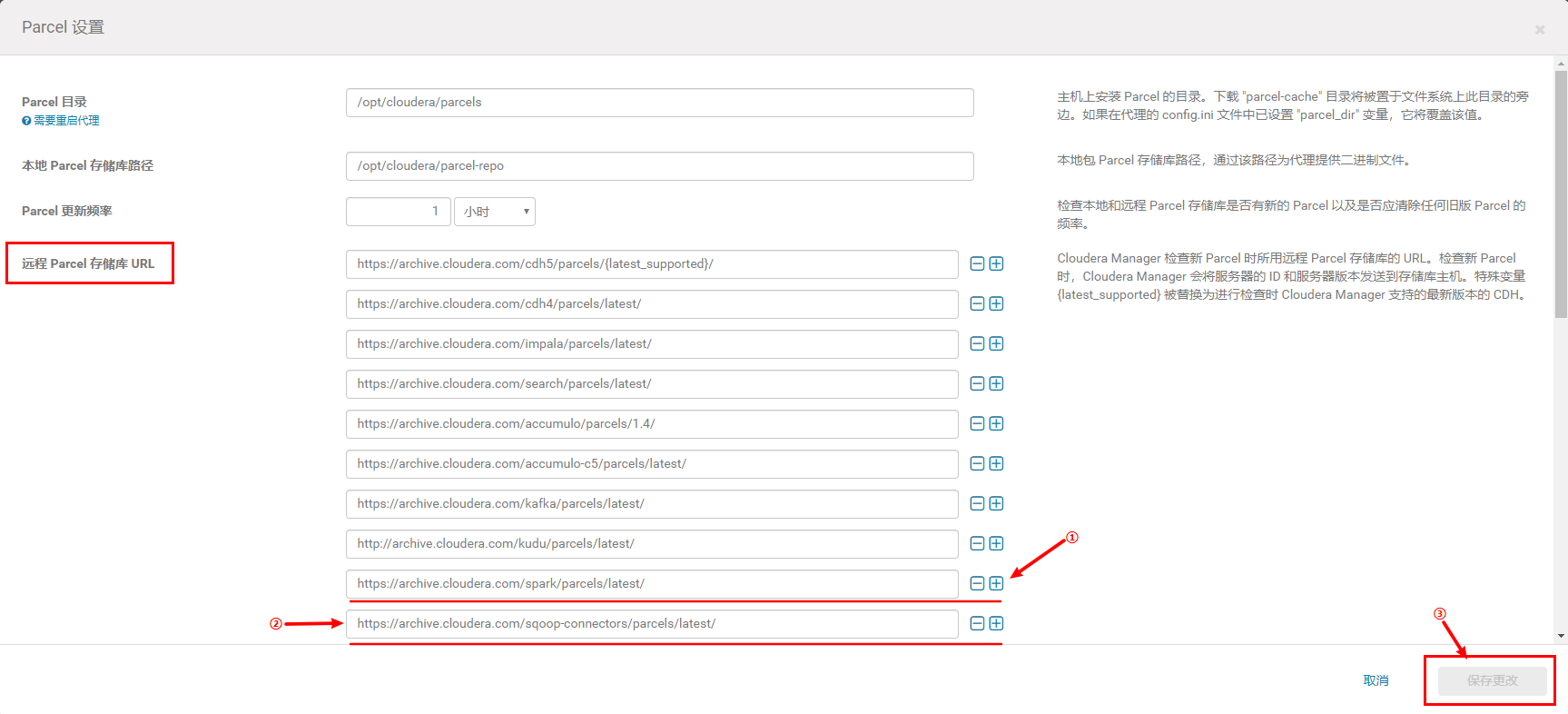
1）点击主机，在下拉菜单中点击Parcel



2）点击配置



3）找到远程Parcel存储库URL，点击最后一栏的加号，增加一栏，输入[http://archive.cloudera.com/gplextras/parcels/latest/](http://archive-primary.cloudera.com/gplextras/parcels/latest/" \t "_blank)，之后点击保存更改

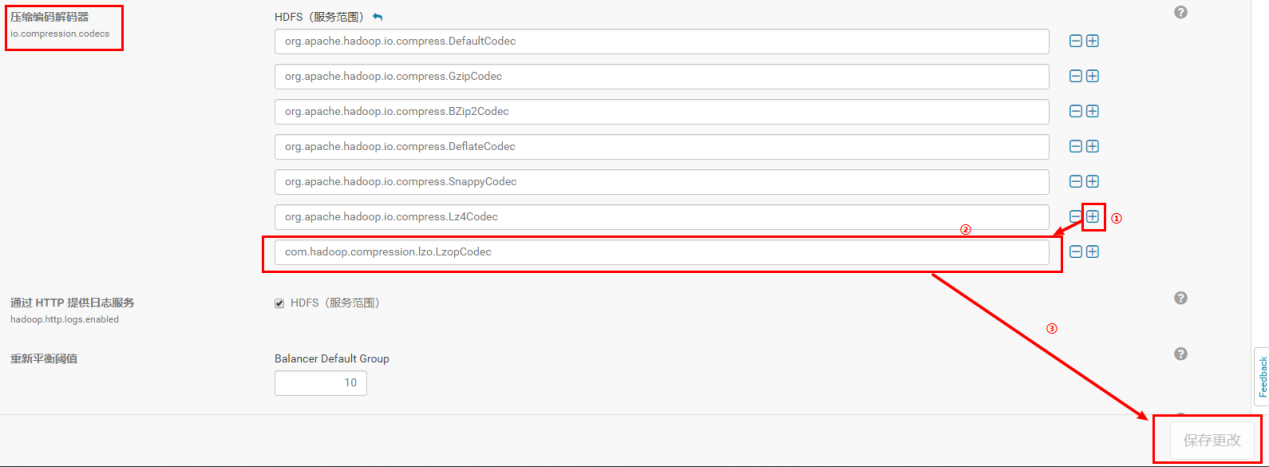


4）返回Parcel列表，可以看到多出了LZO，选择下载，下载完成后选择分配，分配完成后选择激活。

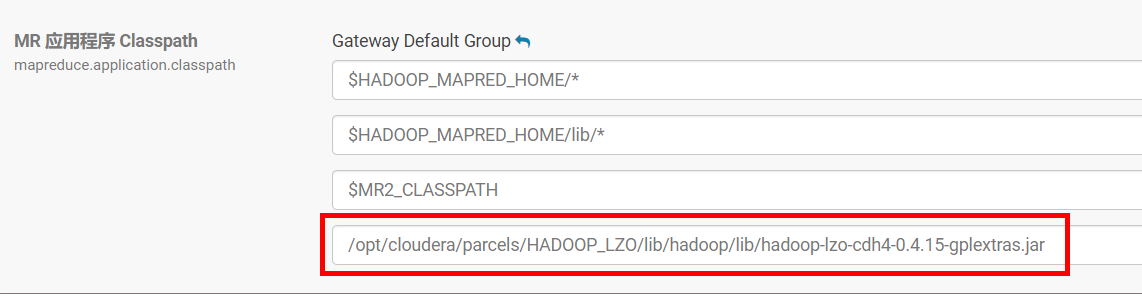


5）安装完LZO后，打开HDFS配置，找到“压缩编码解码器”一栏，点击加号，添加com.hadoop.compression.lzo.LzopCodec后保存更改





6）打开YARN配置，找到MR 应用程序 Classpath，添加/opt/cloudera/parcels/HADOOP\_LZO/lib/hadoop/lib/hadoop-lzo-cdh4-0.4.15-gplextras.jar

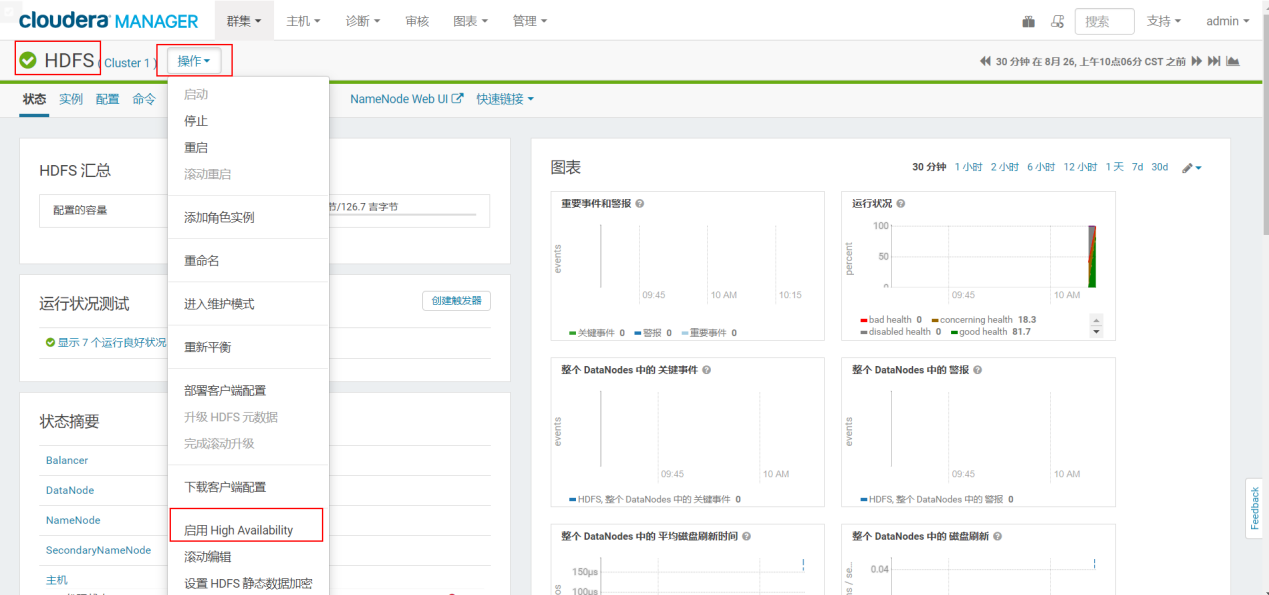


7）更新过期配置，重启进程



2.1.8 配置NameNode HA

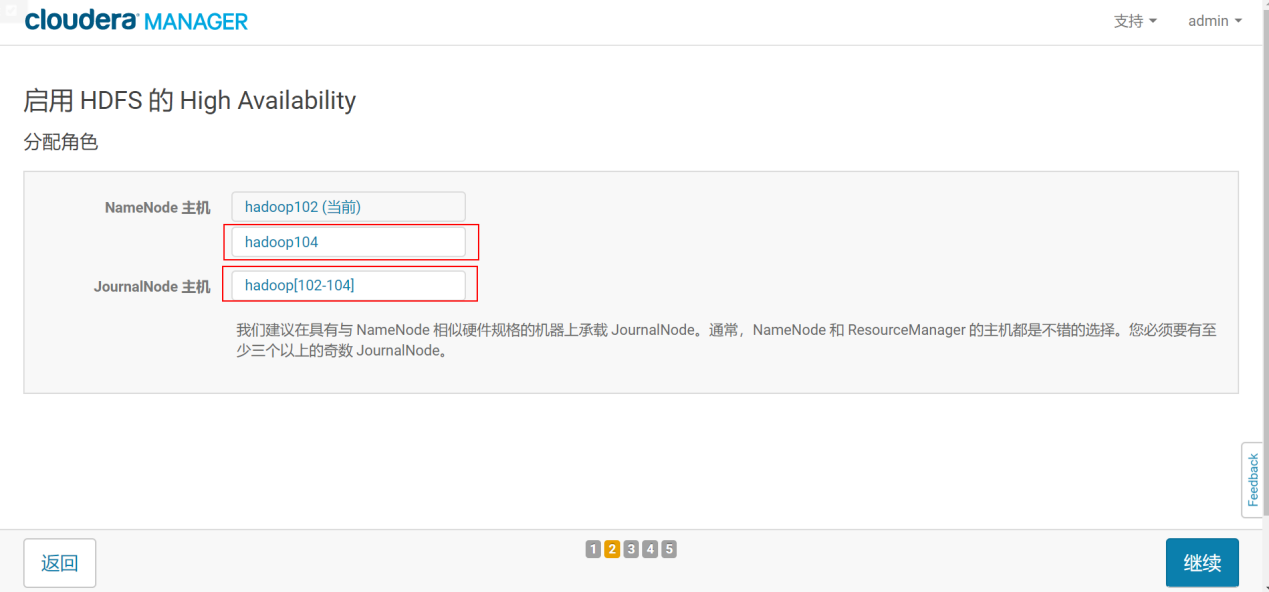
1）进入HDFS页面点击启用High Availability



2）命名



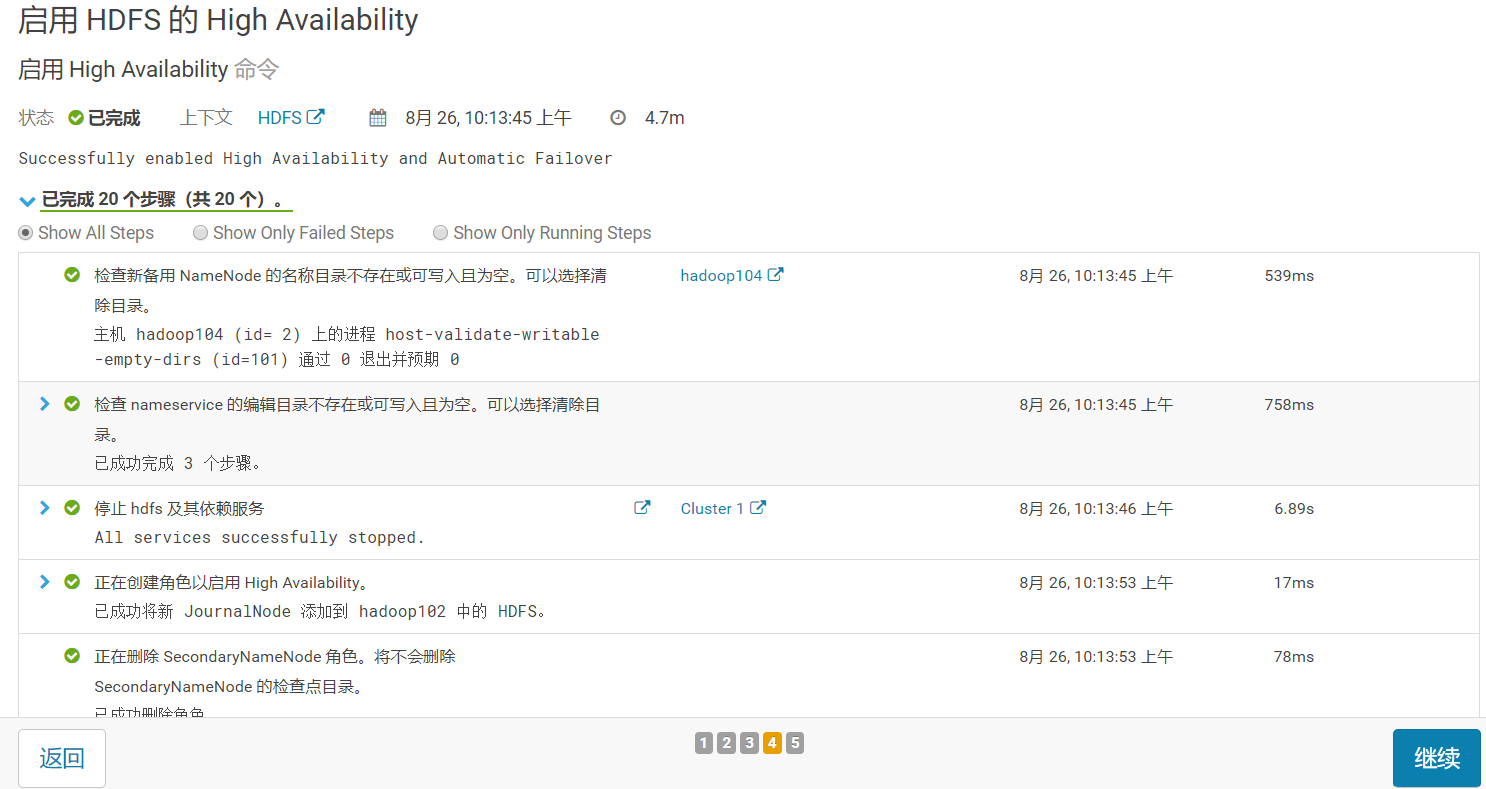
3）分配角色



4）审核更改



5）等待启动服务



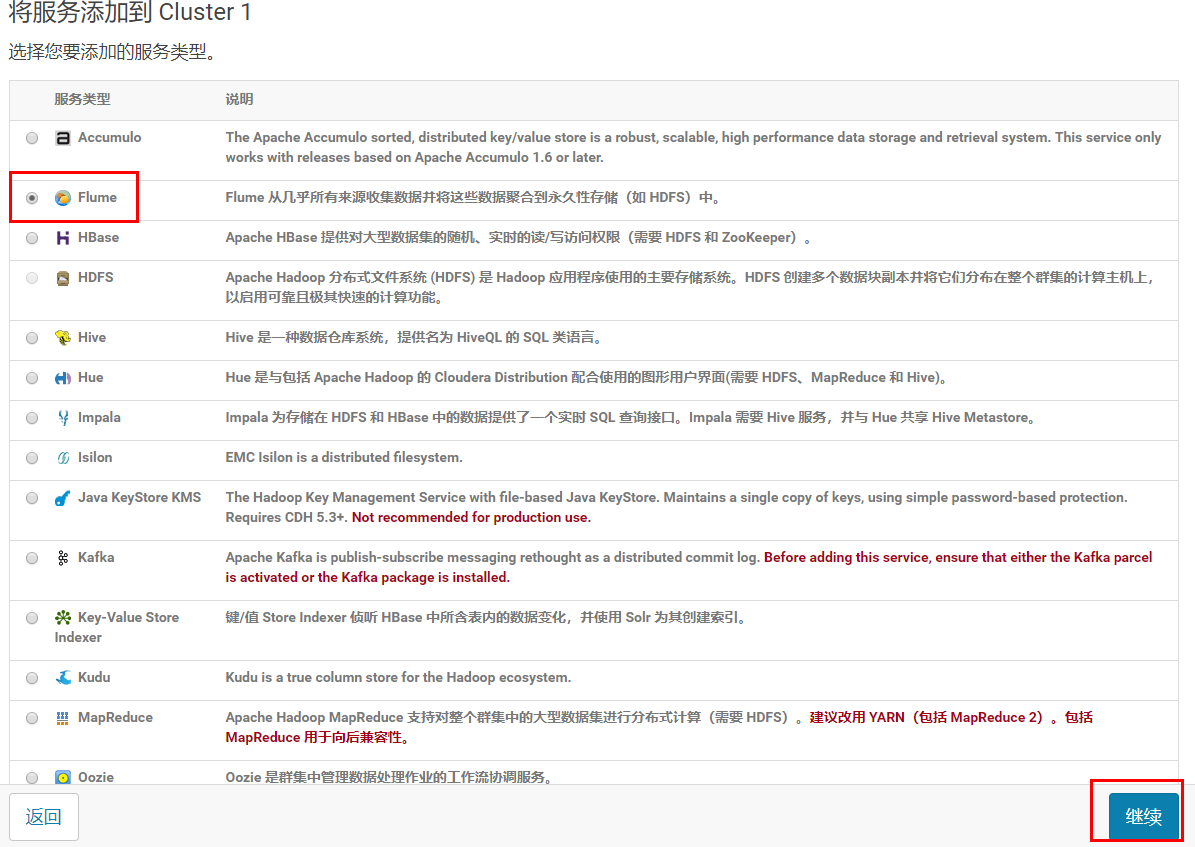
## 2.2 Flume安装

### 2.2.1 日志采集Flume安装

1）添加服务



2）选择Flume，点击继续



3）选择节点



4）完成



1. flume的配置文件在

/opt/cloudera/parcels/CDH/lib/flume-ng/conf目录

1. flume的拦截器的jar包放置在

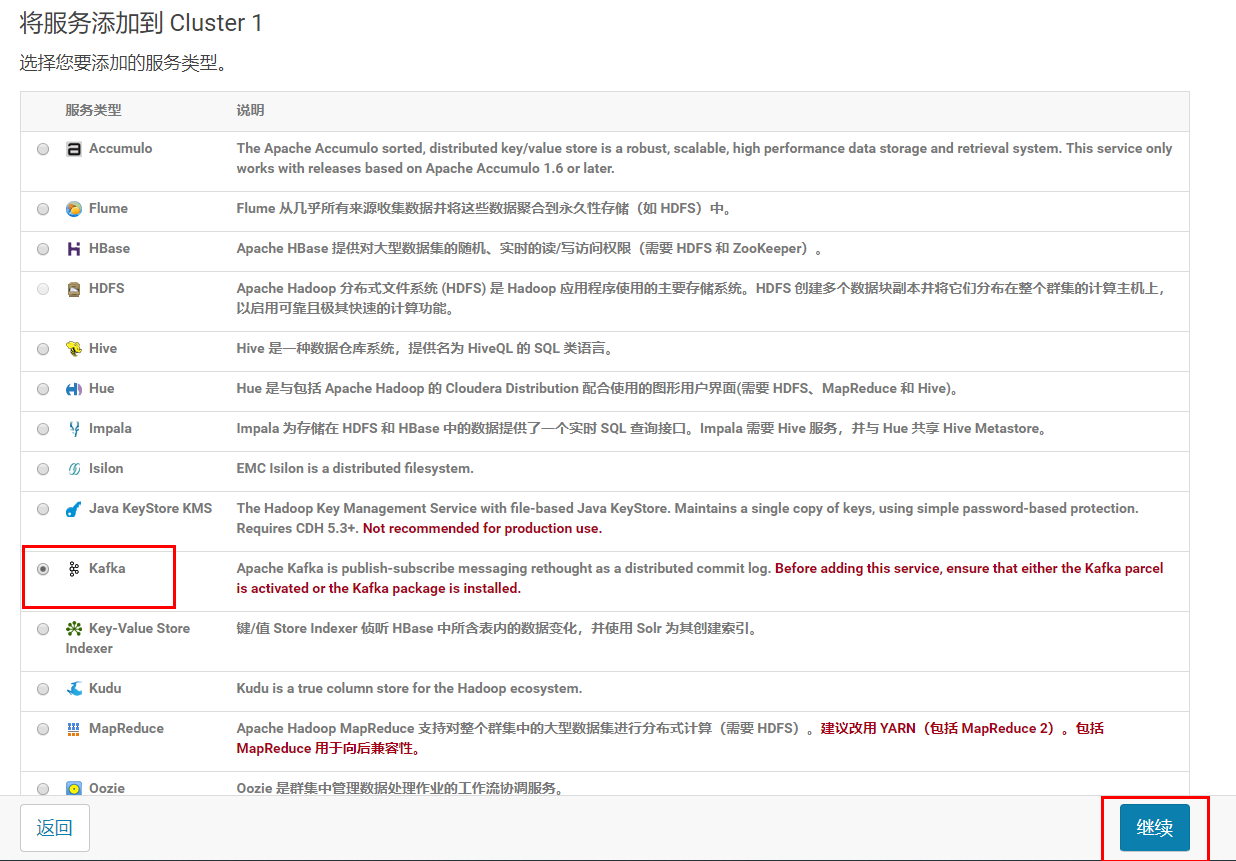
/opt/cloudera/parcels/CDH/lib/flume-ng/lib

2.3 Kafka安装

1) 回到首页，点击添加服务



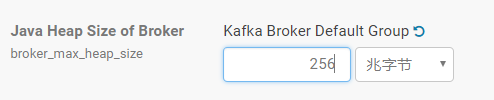
1. 选择Kafka，点击继续



3）Kafka的Broker选择三台机器



4）修改Kafka的堆大小为256M



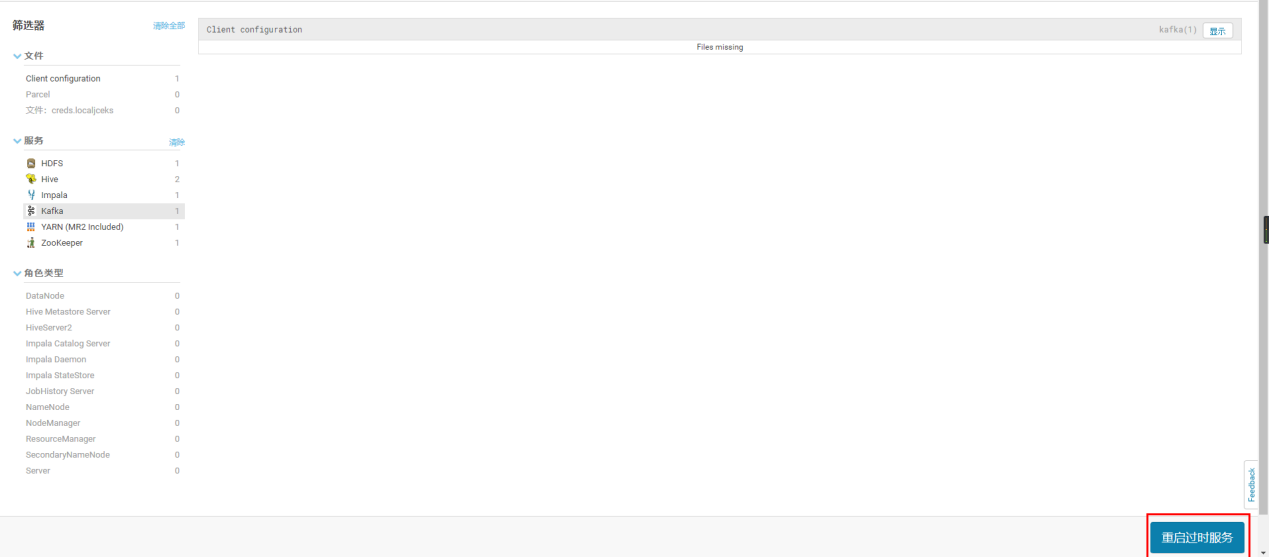
1. 完成

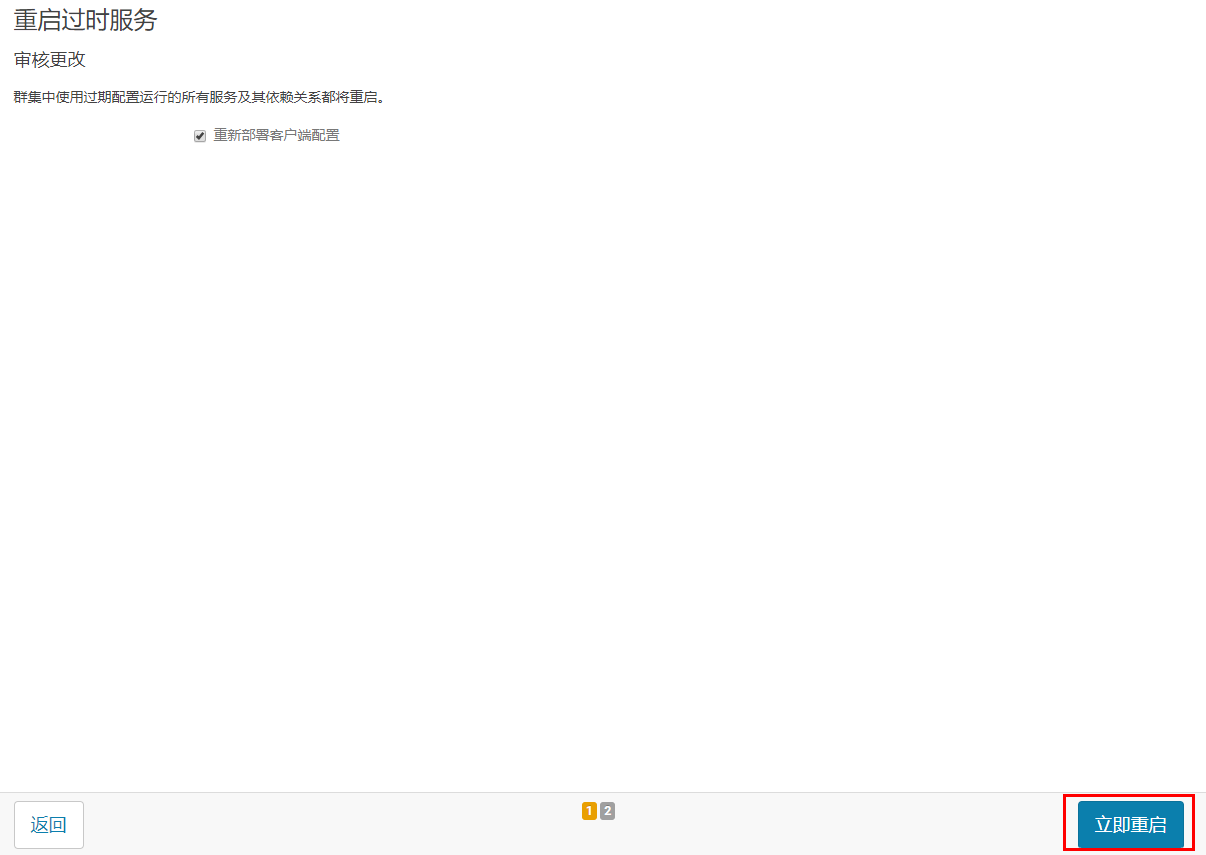


6）重新部署客户端配置

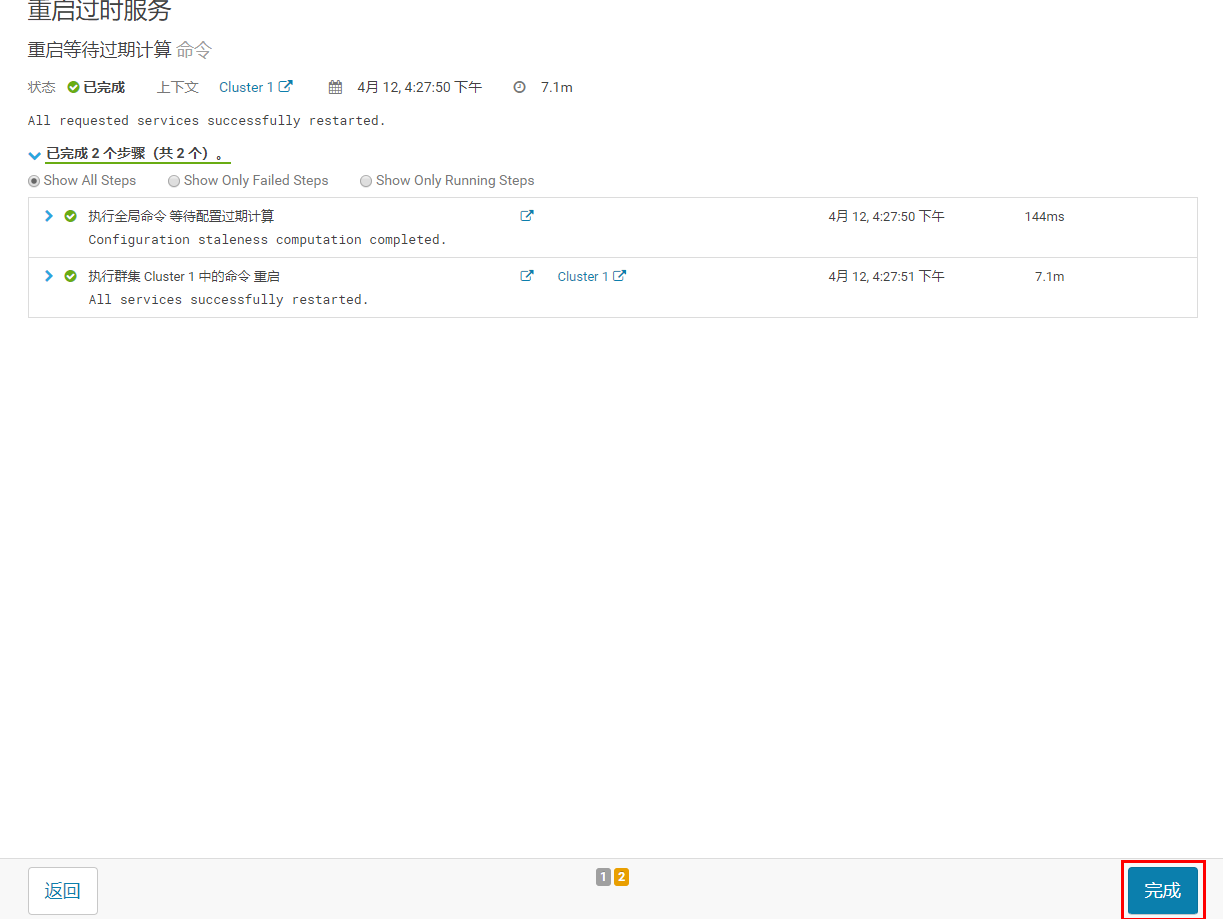


7）点击重启过时服务





1. 等待重启，重启完成后点击完成



2.3.4 查看Kafka Topic

[root@hadoop102 KAFKA]#

/opt/cloudera/parcels/KAFKA/bin/kafka-topics --zookeeper hadoop102:2181 --list

2.2.5 创建 Kafka Topic

进入到/opt/cloudera/parcels/KAFKA目录下分别创建：启动日志主题、事件日志主题。

1）创建启动日志主题

[root@hadoop102 KAFKA]$ bin/kafka-topics --zookeeper hadoop102:2181,hadoop103:2181,hadoop104:2181 --create --replication-factor 1 --partitions 1 --topic topic\_start

2）创建事件日志主题

[root@hadoop102 KAFKA]$ bin/kafka-topics --zookeeper hadoop102:2181,hadoop103:2181,hadoop104:2181 --create --replication-factor 1 --partitions 1 --topic topic\_event

### 2.3.6 删除 Kafka Topic

1）删除启动日志主题

[root@hadoop102 KAFKA]$ bin/kafka-topics --delete --zookeeper hadoop102:2181,hadoop103:2181,hadoop104:2181 --topic topic\_start

2）删除事件日志主题

[root@hadoop102 KAFKA]$ bin/kafka-topics --delete --zookeeper hadoop102:2181,hadoop103:2181,hadoop104:2181 --topic topic\_event

### 2.3.7 生产消息

[root@hadoop102 KAFKA]$

bin/kafka-console-producer.sh --broker-list hadoop102:9092 --topic topic\_start

>hello world

>atguigu atguigu

### 2.3.8 消费消息

[root@hadoop103 KAFKA]$

bin/kafka-console-consumer.sh \

--bootstrap-server hadoop102:9092 --from-beginning --topic topic\_start

--from-beginning：会把first主题中以往所有的数据都读取出来。根据业务场景选择是否增加该配置。

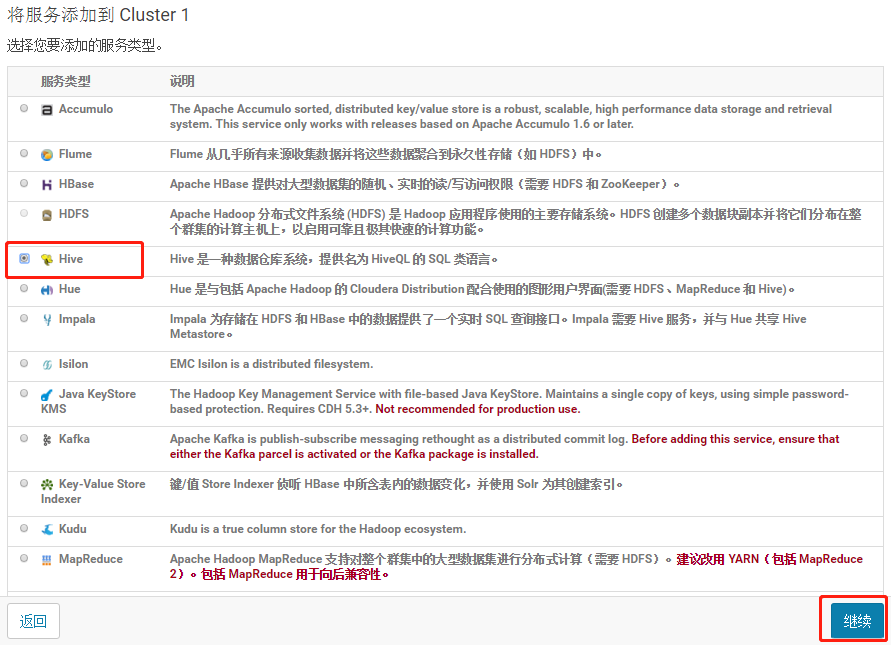
# 数仓之数仓搭建环境准备

## 3.1 Hive安装

3.1.1 添加服务

# 

3.1.2 添加Hive服务



3.1.3 将 Hive 服务添加到 Cluster 1

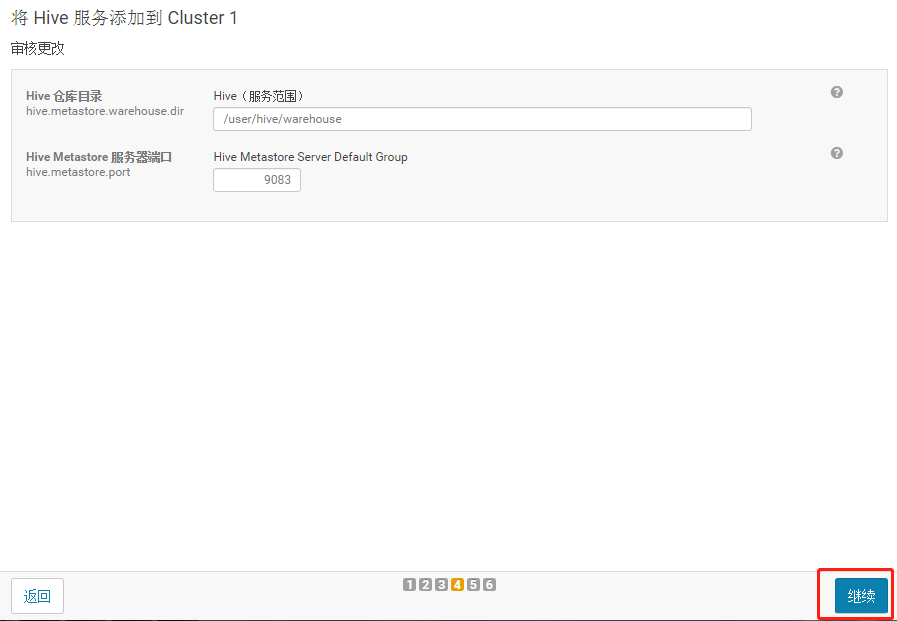


3.1.4 配置hive元数据



3.1.5 测试通过后继续





3.1.6 自动启动Hive进程

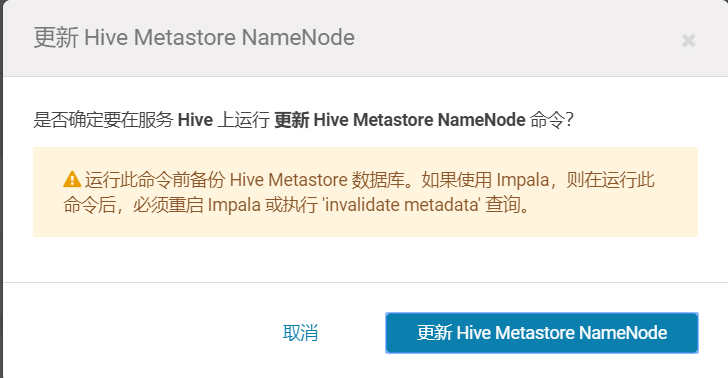




3.1.7 修改Hive配置

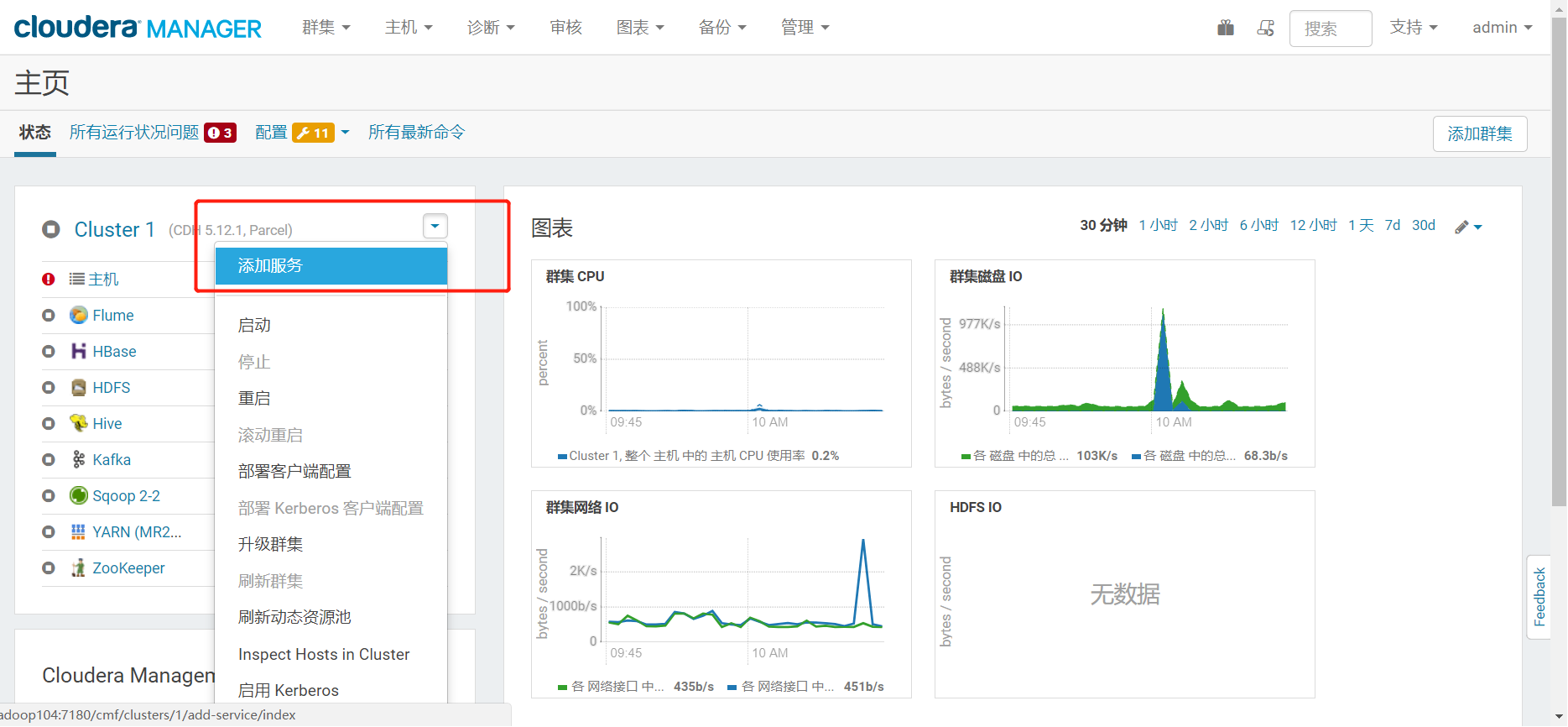


/opt/cloudera/parcels/HADOOP\_LZO/lib/hadoop/lib/

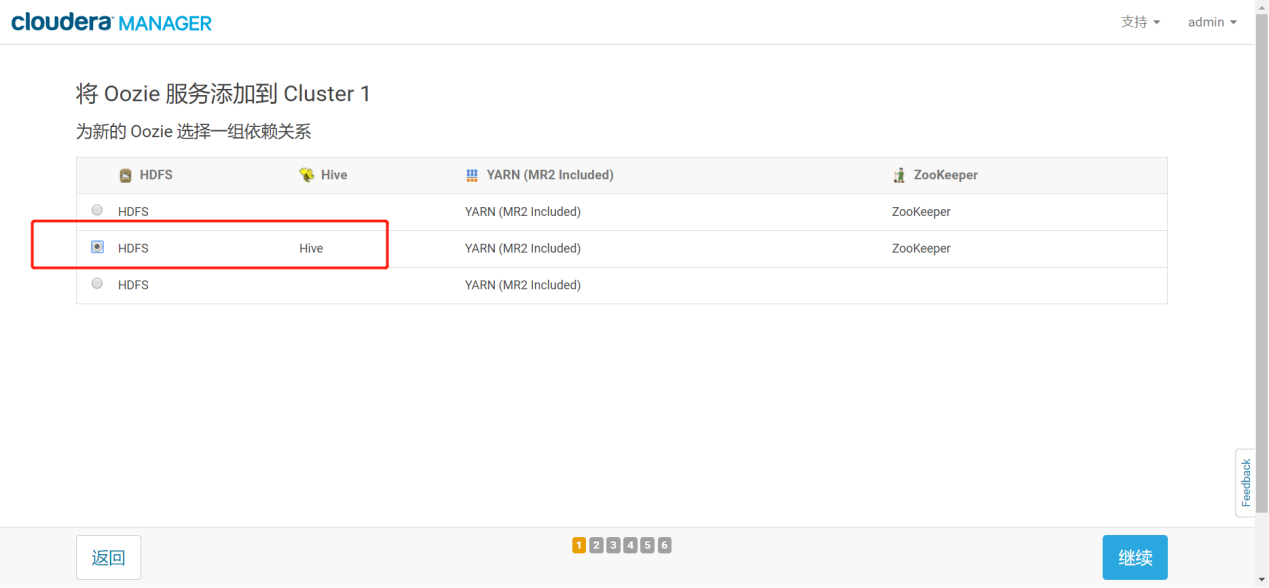
3.1.8 更新Hive Metastore NameNode（先停止Hive）

## 3.2 Oozie安装

### 3.2.1 添加Oozie服务



### 3.2.2 选择集群节点



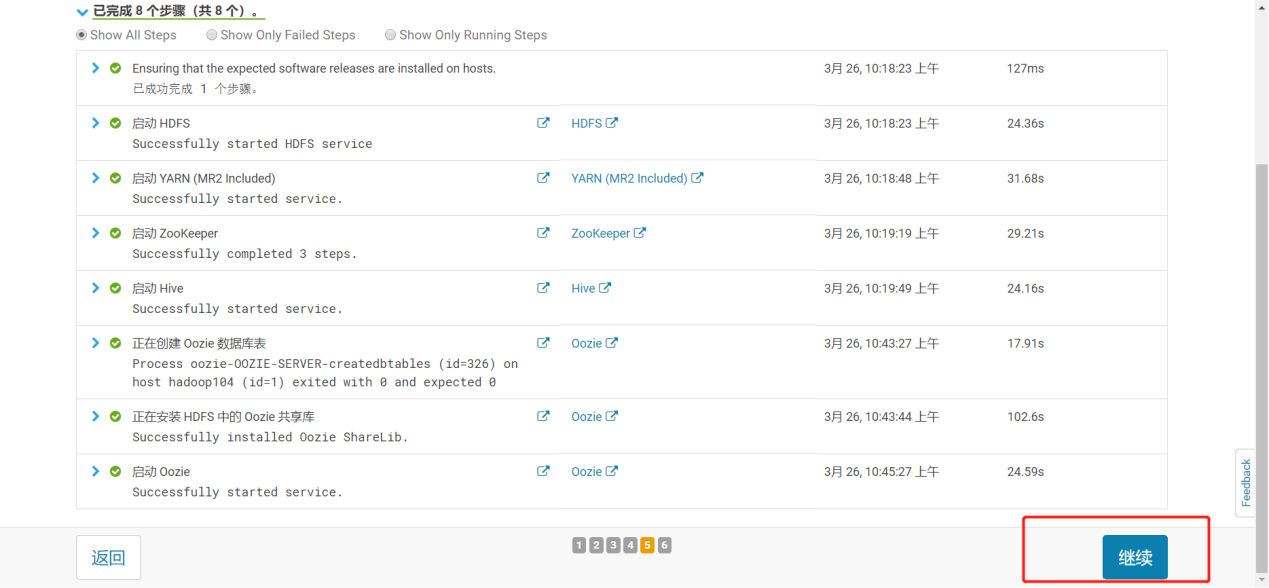
### 3.2.3 选择有MySQL的节点安装



### 3.2.4 链接数据库



### 3.2.5 一路继续到完成



## 3.3 Hue安装

3.3.1 Hue概述

1）Hue来源

HUE=Hadoop User Experience（Hadoop用户体验），直白来说就一个开源的Apache Hadoop UI系统，由Cloudera Desktop演化而来，最后Cloudera公司将其贡献给Apache基金会的Hadoop社区，它是基于Python Web框架Django实现的。通过使用HUE我们可以在浏览器端的Web控制台上与Hadoop集群进行交互来分析处理数据。

2）Hue官网及使用者

官网网站：<http://gethue.com/>

3.3.2 安装前的准备

1）在LoadBalancer节点安装mod\_ssl

[root@hadoop102 ~]# yum -y install mod\_ssl

2）查看/usr/lib64/mysql下有没有libmysqlclient\_r.so.16，如果没有，上传libmysqlclient\_r.so.16到/usr/lib64/mysql，并软链接到/usr/lib64/

[root@hadoop102 ~]# ls /usr/lib64/mysql

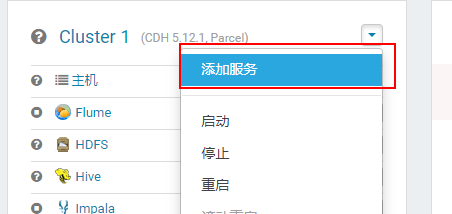
[root@hadoop103 mysql]#

scp /usr/lib64/mysql/libmysqlclient\_r.so.16 root@hadoop102:/usr/lib64/mysql/

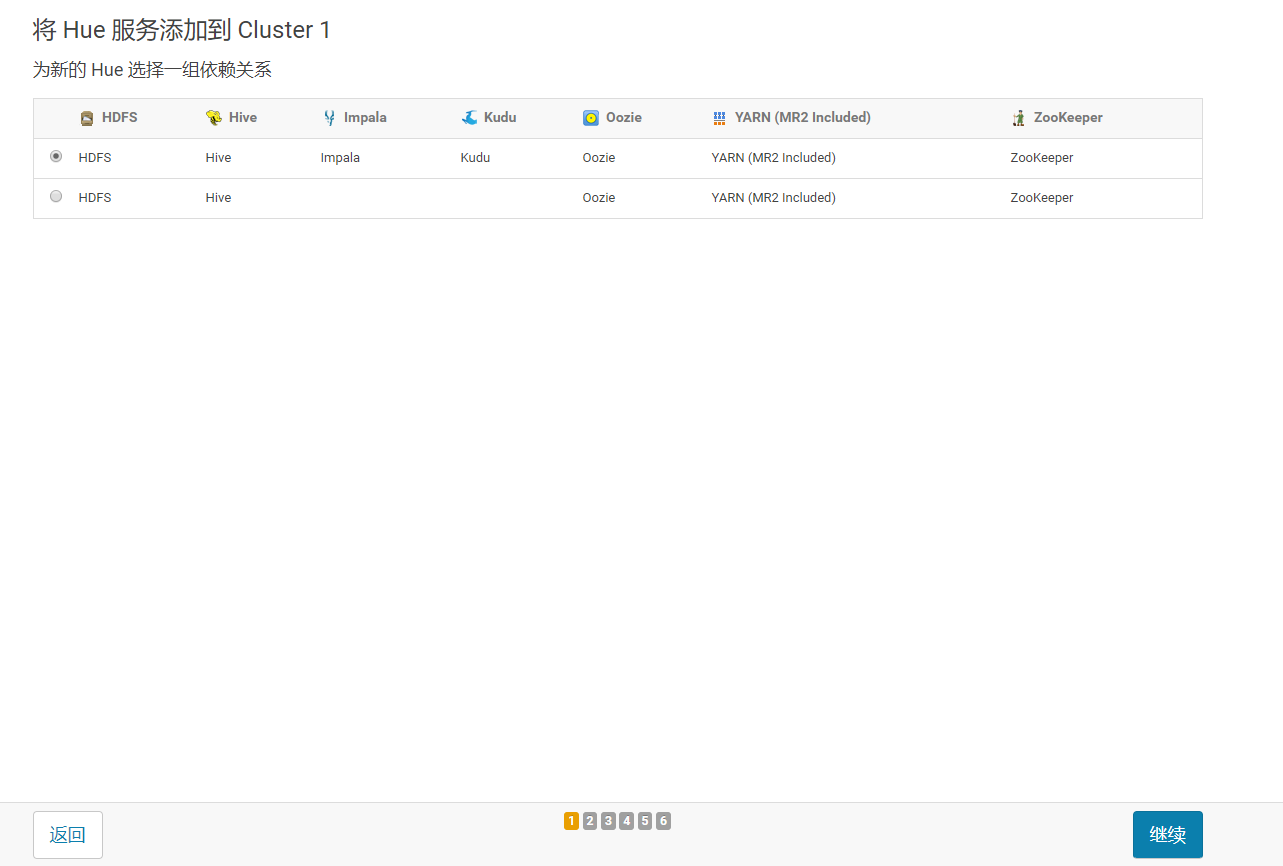
[root@hadoop102 ~]ln -s /usr/lib64/mysql/libmysqlclient\_r.so.16 /usr/lib64/libmysqlclient\_r.so.16

3.3.3 HUE安装步骤

1）添加HUE服务



2）选择集群节点



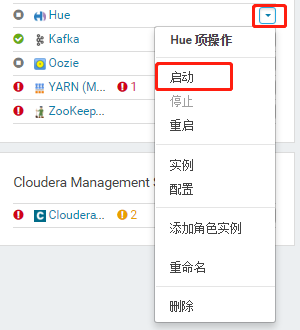
3）分配角色

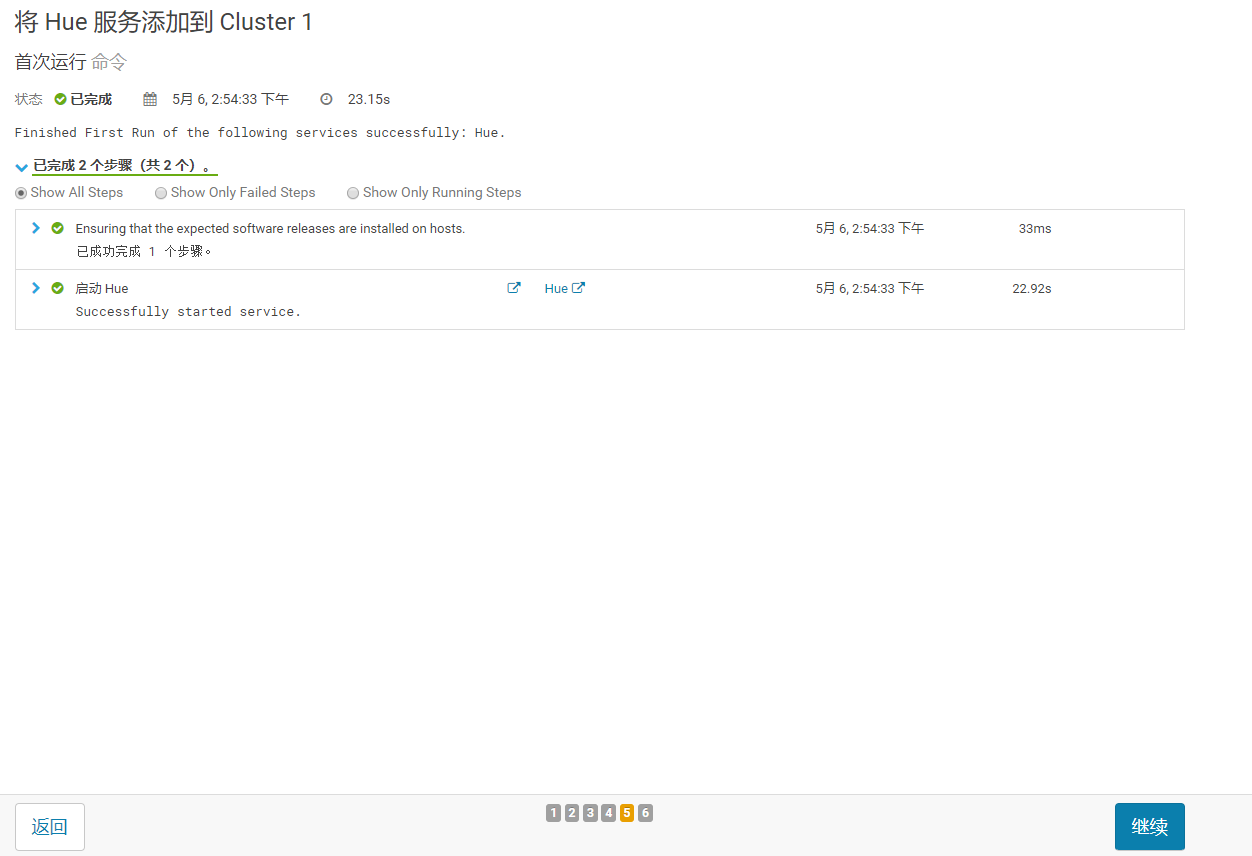


4）配置数据库



5）安装完成





6）HUE页面

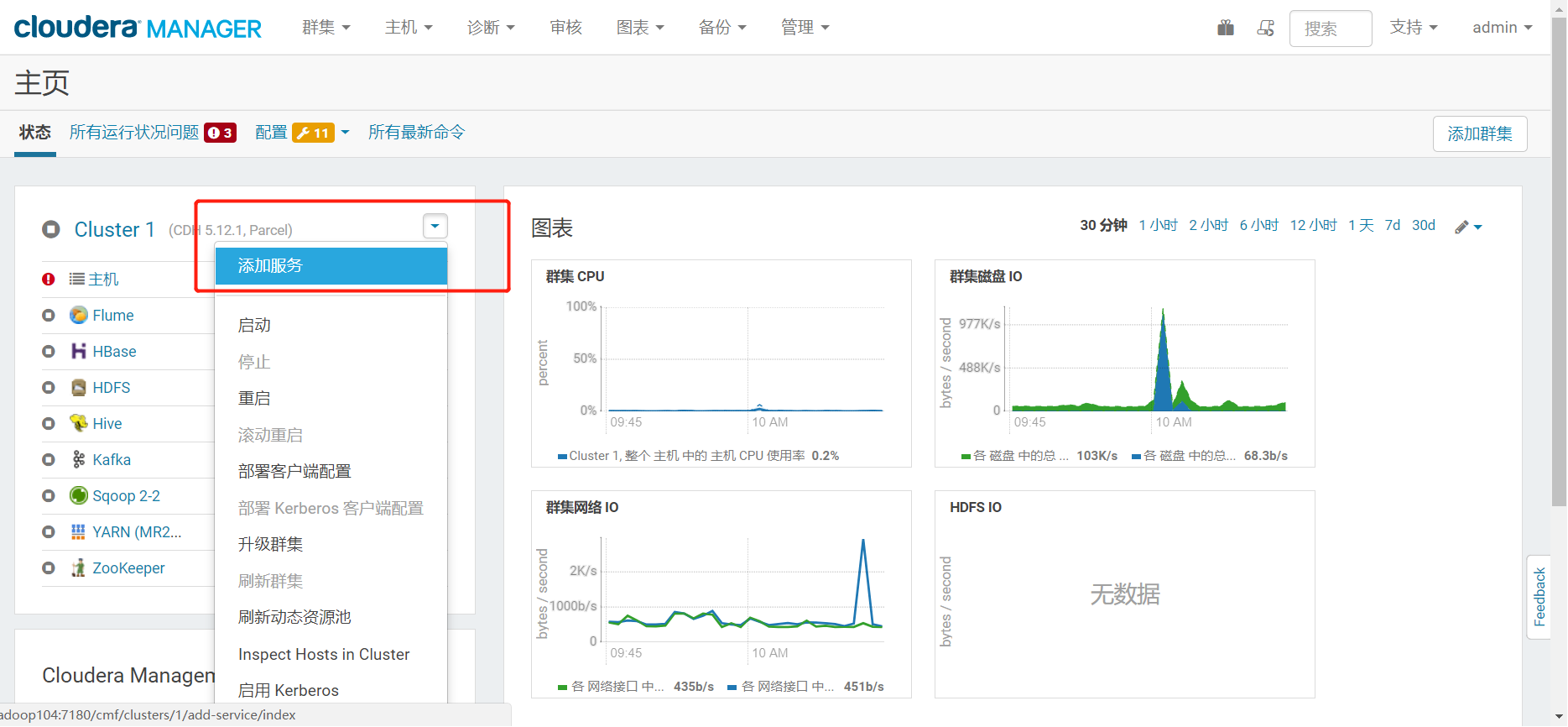
<http://hadoop102:8888>（未优化）或<http://hadoop102:8889>（优化）

第一次开启HUE会出现以下页面，此时输入的用户名和密码可以随意，之后登录页面以第一次输入的账号密码为依据。例如，用户名：admin 密码：admin



### 3.4 Sqoop安装

1）添加服务



2）选择Sqoop



目前选择sqoop1即可

3）选择节点





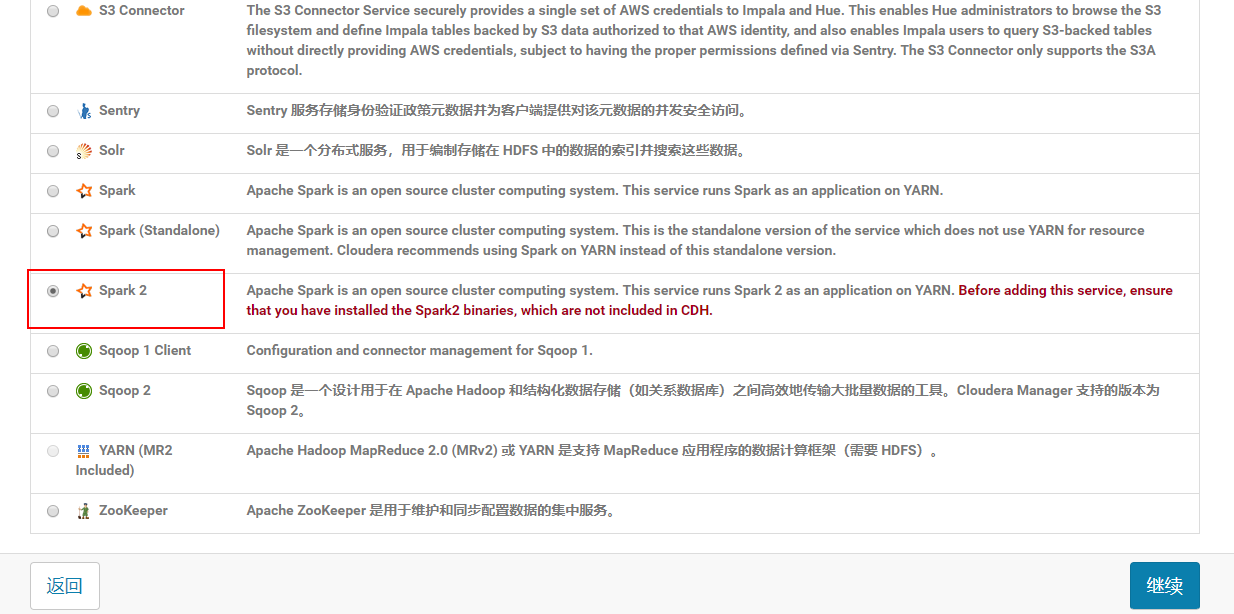
### 3.5 spark安装

Spark2和spark是一样的东西,spark2只是版本升级了,咱们系统中的CDH6.3.1不需要升级,所以部署的时候选择spark就行,不需要升级,下面的部署过程,用的升级的spark2

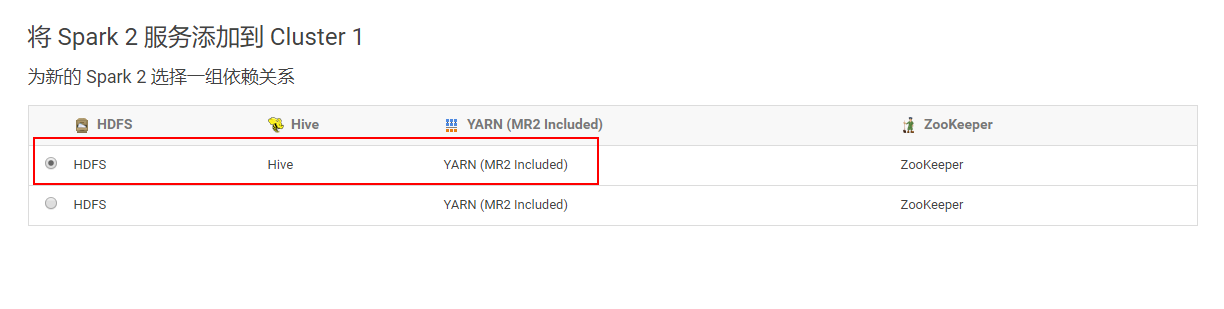
1) 添加服务



### 点击Spark2继续



### 选择一组依赖关系



### 角色分配



### 部署并启动

