使用 Cloudera 部署，管理 Hadoop 集群

1. Cloudera 介绍

Hadoop 是一个开源项目，Cloudera 对 Hadoop 进行了商业化，简化了安装过程，并对 hadoop 做了一些封装。

根据使用的需要，Hadoop 集群要安装很多的组件，一个一个安装配置起来比较麻烦，还要考虑 HA，监控等。

使用 Cloudera 可以很简单的部署集群，安装需要的组件，并且可以监控和管理集群。

CDH 是 Cloudera 公司的发行版，包含 Hadoop，Spark，Hive，Hbase 和一些工具等。

Cloudera 有两个版本：

Cloudera Express 版本是免费的 Cloudera Enterprise (60 天试用期)需要购买注册码

2. 安装 Cloudrea Manager，部署 Hadoop 集群

2.1 安装方法

先安装 Cloudrea Manager，再通过 Cloudrea Manager 在节点上安装 Cloudrea Manager 客户端，CDH，管理工具。

官方文档：

https://www.cloudera.com/documentation/manager/5-1-x.html

环境需求：

1. 关闭 selinux

2. 各节点可以 SSH 登陆

3. 在/etc/hosts 中添加各节点的主机名

2.2 安装 Cloudrea Manager

可以通过官方的一键安装包，也可以通过 yum 或 rpm 安装。

下面介绍用官方的一键安装包安装。

本次安装环境为 CnetOS 7，在 3 台机器上进行安装

test165 (cloudera manager server)

test166 (cloudera manager agent)

test167 (cloudera manager agent)

2.2.1 下载一键安装包

http://archive.cloudera.com/cm5/installer/latest/

下载最新版： cloudera-manager-installer.bin

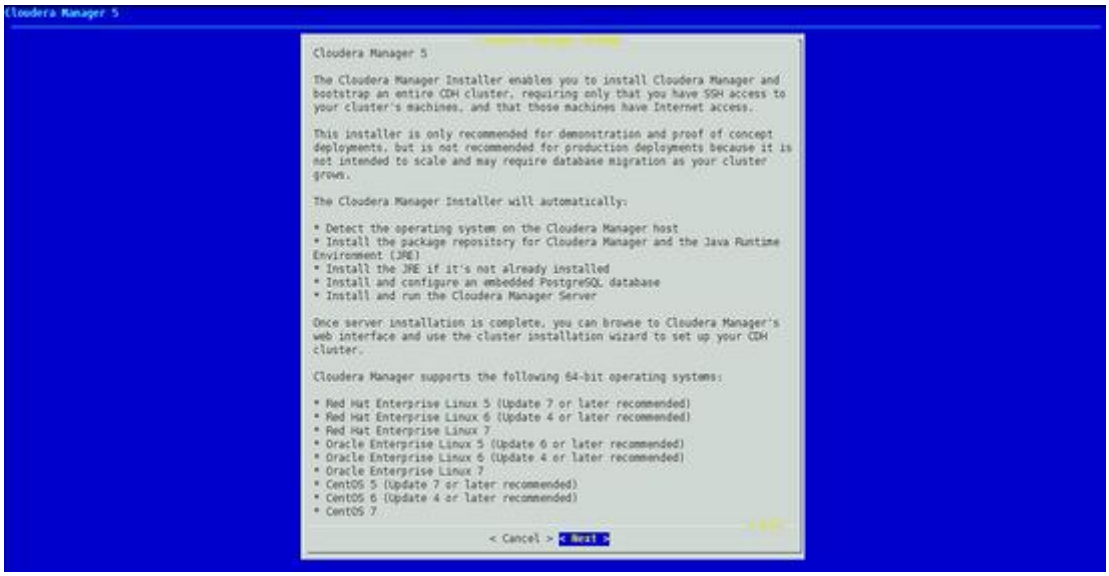
2.2.2 安装 cloudera manager

在 test165 上安装 cloudera manager server，启动安装向导

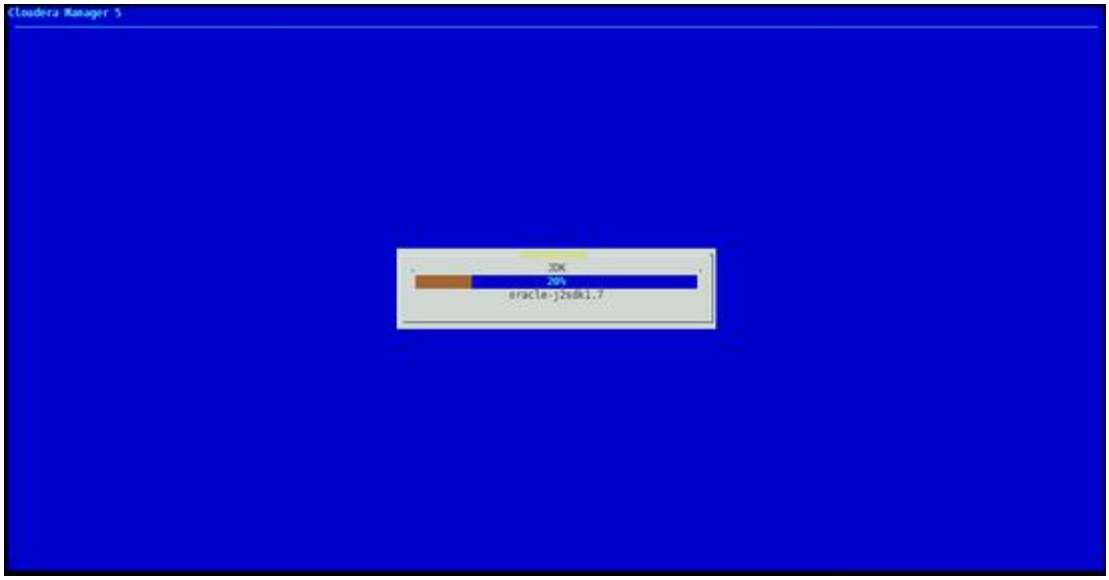
chmod a+x cloudera-manager-installer.bin

./cloudera-manager-installer.bin

出现下面画面

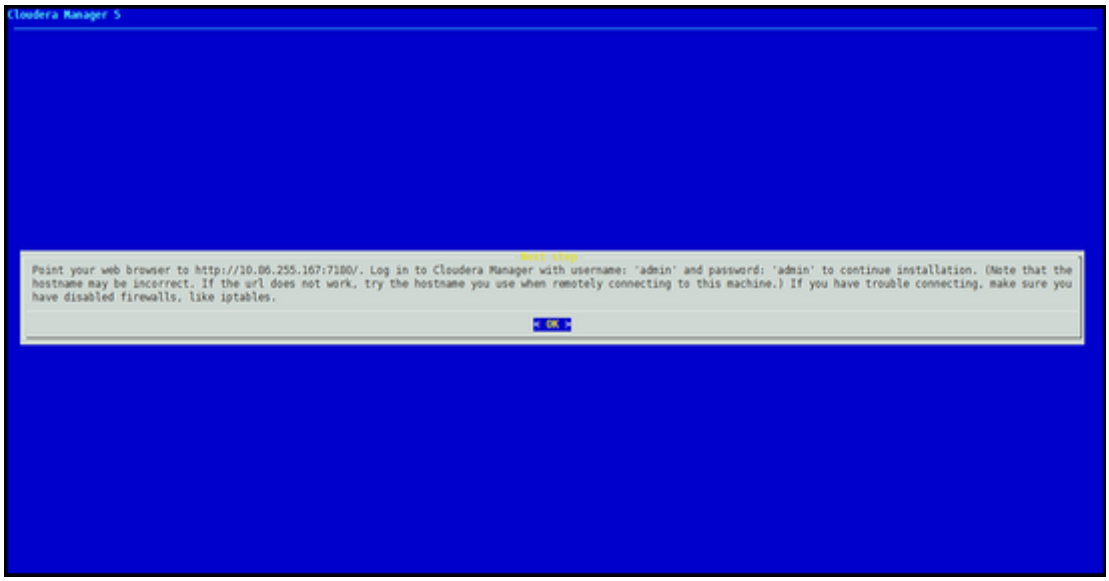


一路选择< Next > 和 < Yes >，开始安装。



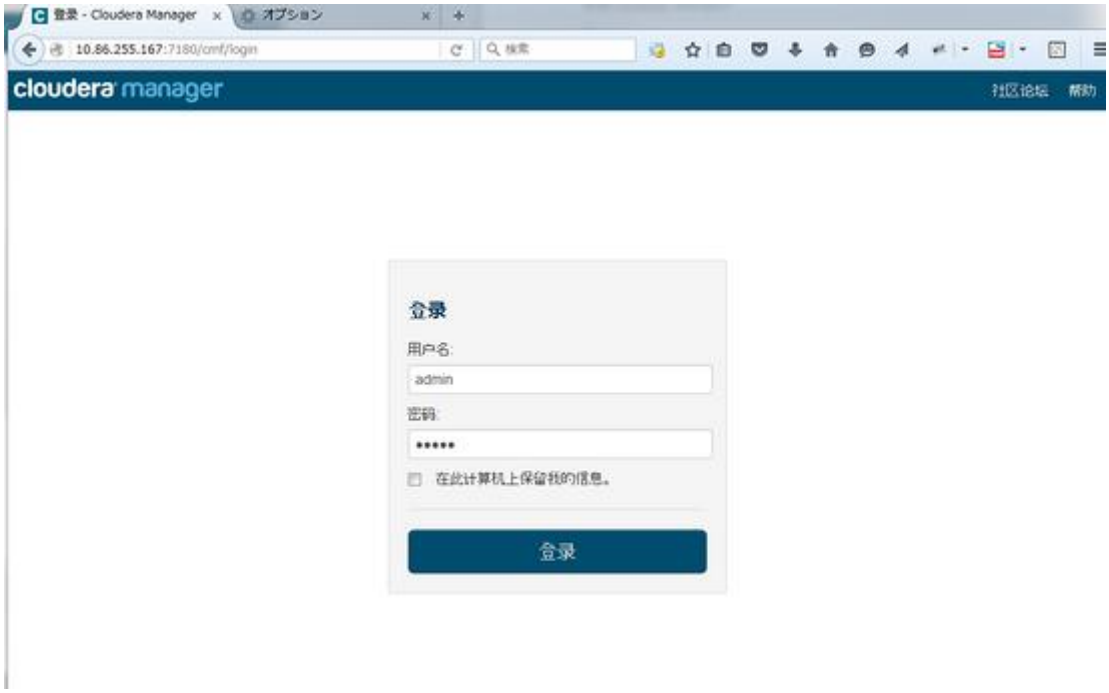
需要下载 JAVA 和 Cloudrea Manager，共 600 多 MB，根据网络情况，会花一些时间。

出现下面页面，安装完成。



安装完成后，访问 Cloudrea Manager 的页面，用户名密码都是 admin

http://IP 或主机名:7180/



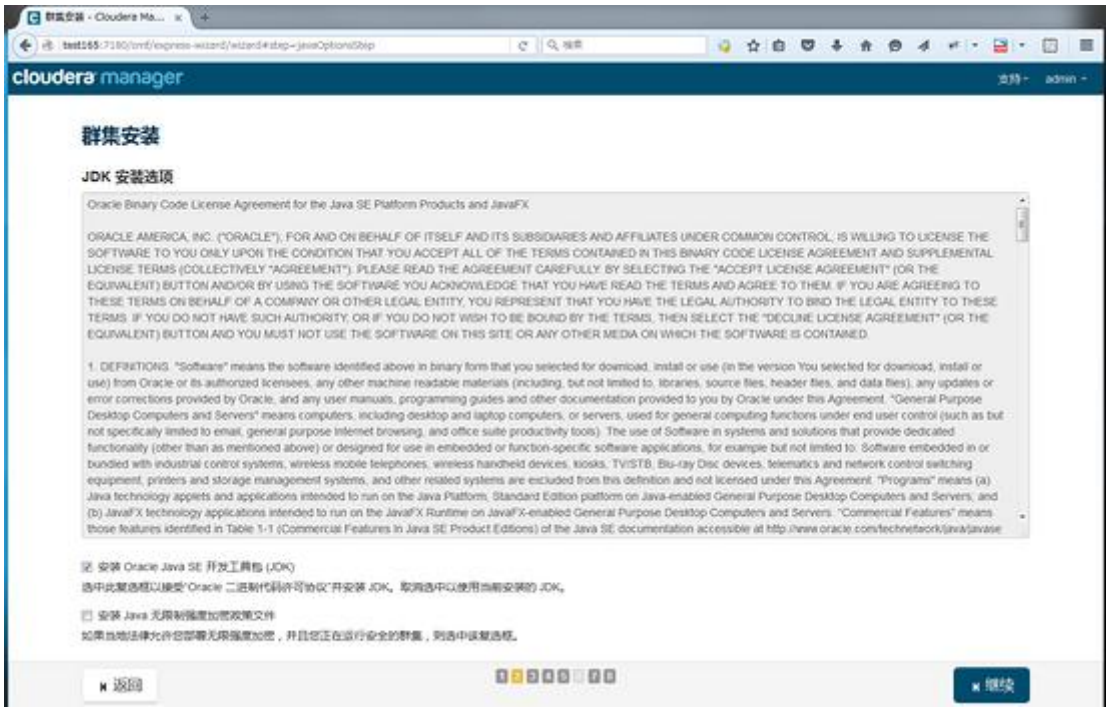
2.2.3 安装 cloudera manager agent

登录 Cloudrea Manager 页面，选择要安装的版本，本次安装的是 Cloudera Express

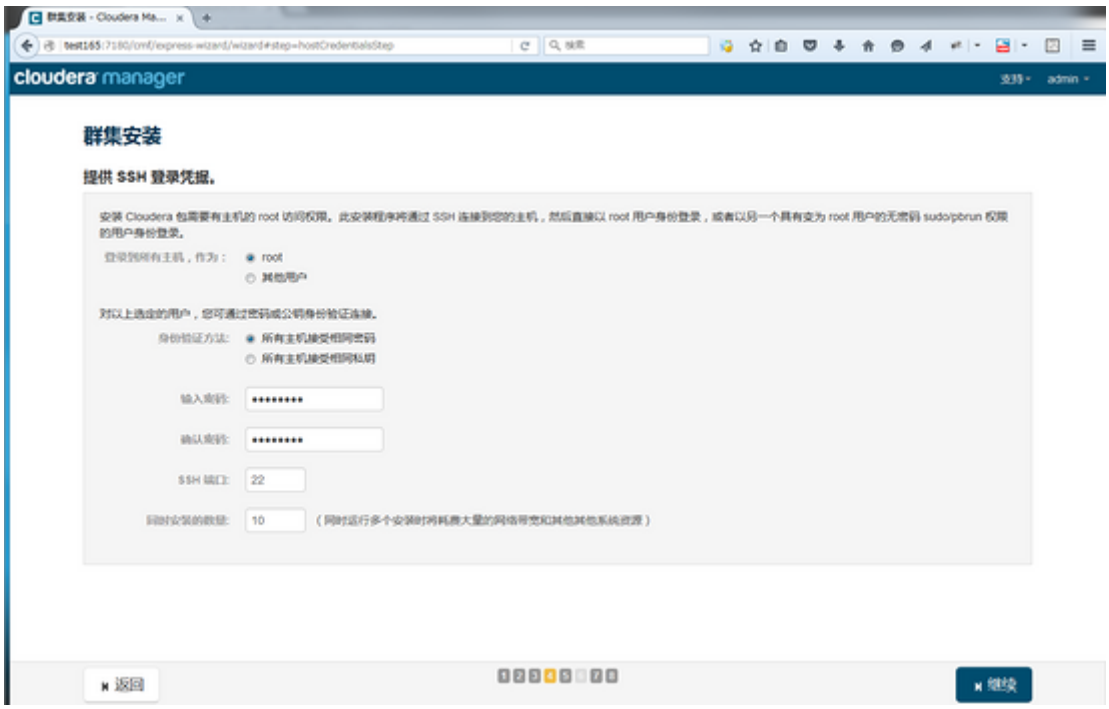
选择要安装 CDH 的主机，用主机名或 IP 搜索，本次是在三个节点上安装 CDH

选择使用 Parcel 安装，选者 CDH 版本

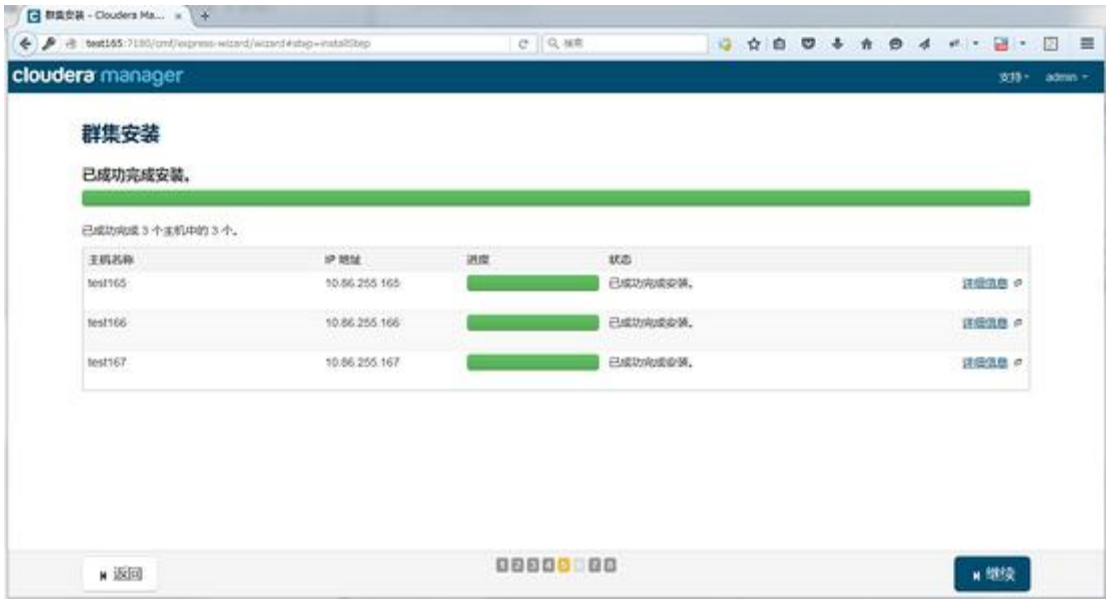
选择安装 JDK



提供 SSH 登录信息



开始安装 JDK 和 cloudera manager agent



如果安装过程中，下载安装 jdk 或 cloudera-manager-agent 失败，可以在节点上手动安装，然后再在 Cloudrea Manager 上继续安装

```
# yum -y install jdk

# yum -y install oracle-j2sdk1.7

# yum -y install cloudera-manager-agent
```

下载 Parcel 并分配 Parcel 到各节点

Parcel 包 1.5G 左右，需要一段时间，为了提高安装速度，可以先把包下载到 Cloudrea Manager 本地，配置本地源

parcel 下载地址：

http://archive.cloudera.com/cdh5/parcels/5.5.1/

将下面文件拷贝到/opt/cloudera/parcel-repo/文件夹下

CDH-5.5.1-1.cdh5.5.1.p0.11-el7.parcel

CDH-5.5.1-1.cdh5.5.1.p0.11-el7.parcel.sha

manifest.json

安装完成后，点继续，到检查结果的页面

检查主机正确性时出现 “Cloudera 建议将 /proc/sys/vm/swappiness 设置为 0。当前设置为 30。” 的警告，进行如下设定

```
# vi /etc/sysctl.conf

vm.swappiness = 0

# sysctl -p
```

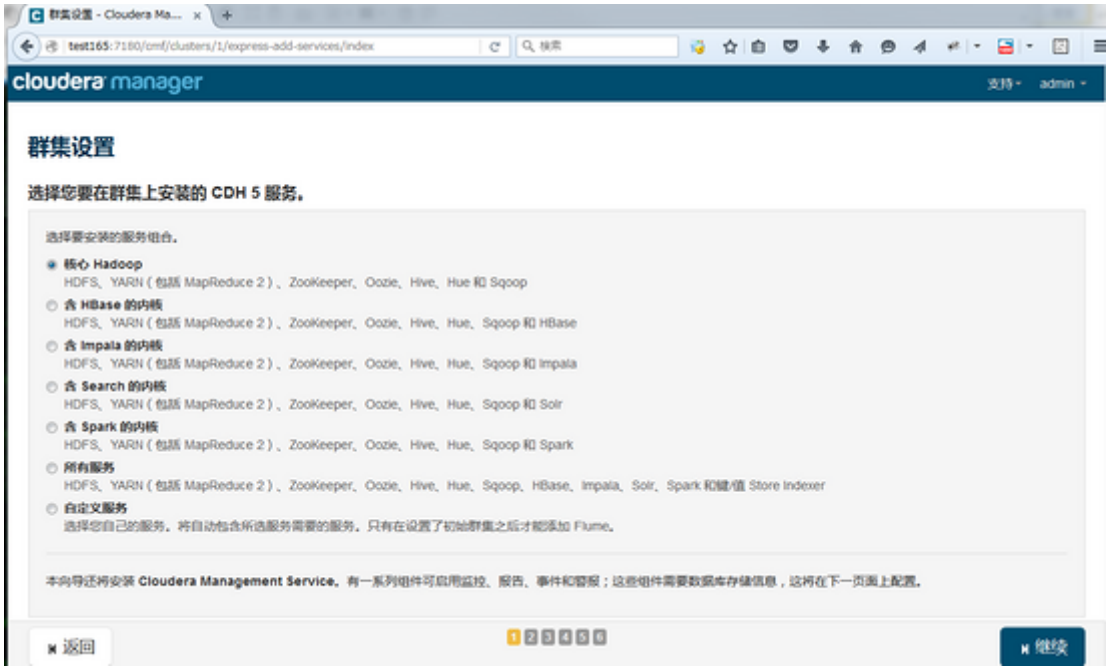
检查主机正确性时出现 “已启用“透明大页面”，它可能会导致重大的性能问题。” 的警告，进行如下设定

```
echo never > /sys/kernel/mm/transparent_hugepage/enabled

echo never > /sys/kernel/mm/transparent_hugepage/defrag

# vi /etc/rc.local

echo never > /sys/kernel/mm/transparent_hugepage/enabled

echo never > /sys/kernel/mm/transparent_hugepage/defrag
```

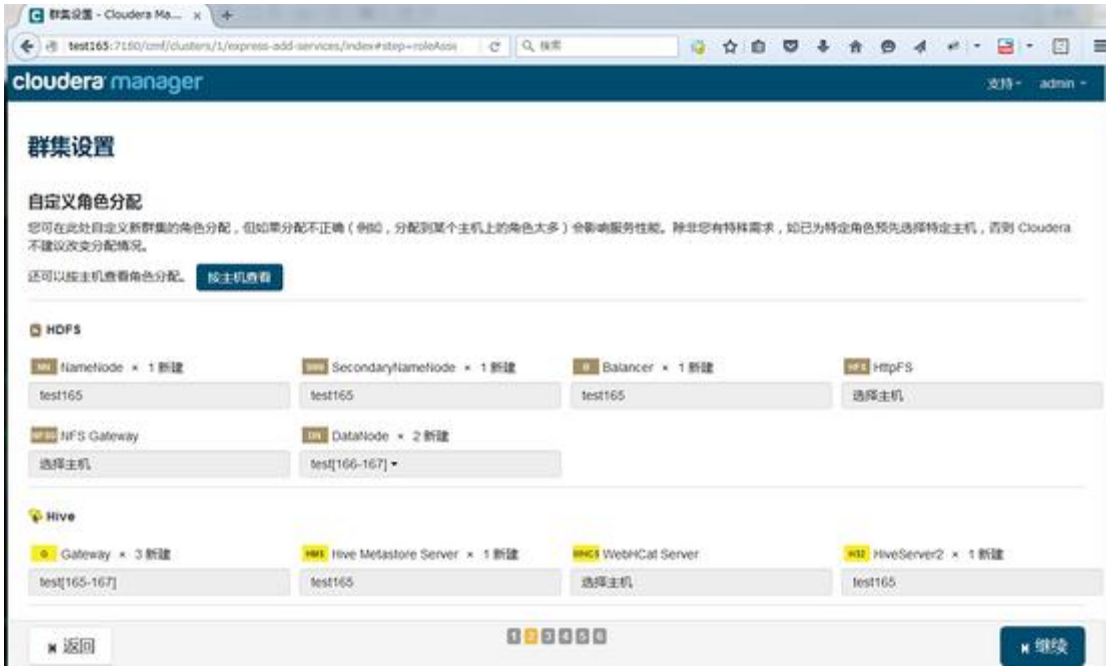
2.3 安装集群，包括 Hadoop，YARN，Hive 等

检查主机正确性后，点击完成，进入集群配置

选择要安装的服务，可以选择组合或自定义

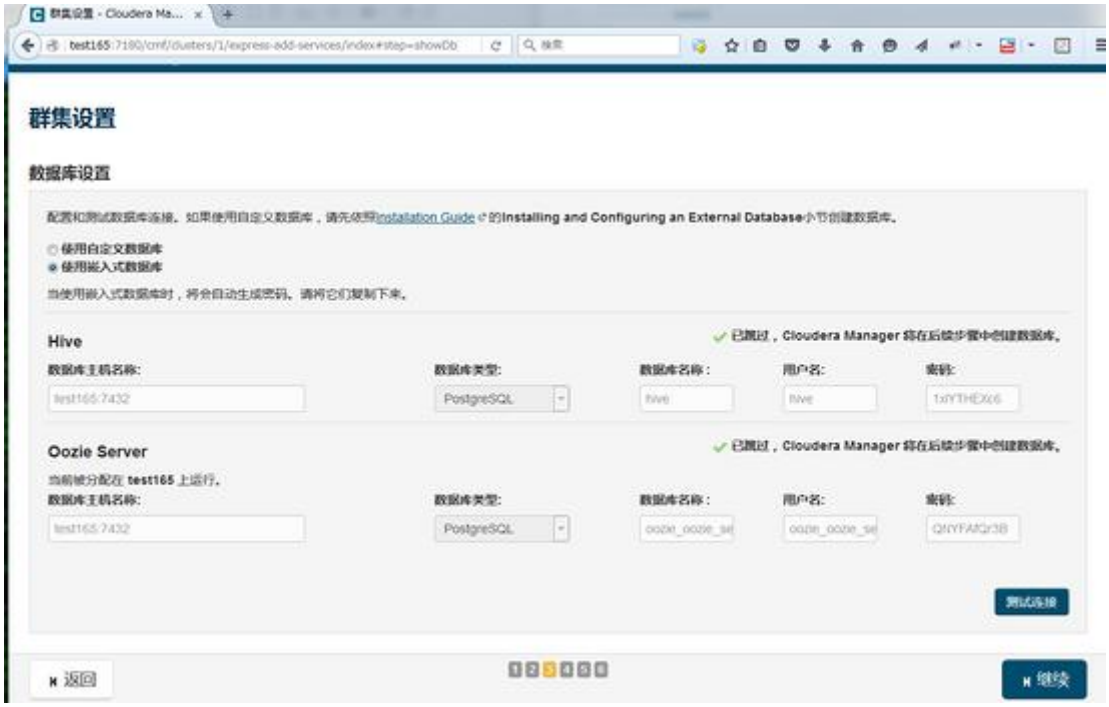


配置各节点间如何分配

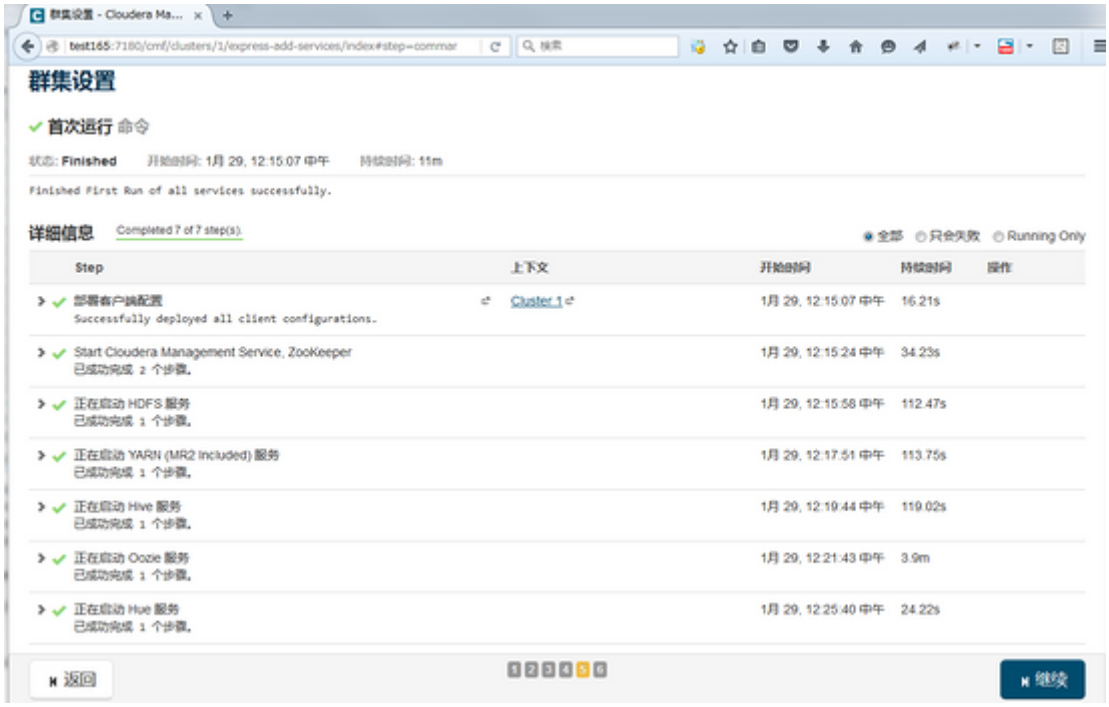


注意： HDFS 的 Data Node 最少 3 个。

测试数据库连接



开始安装

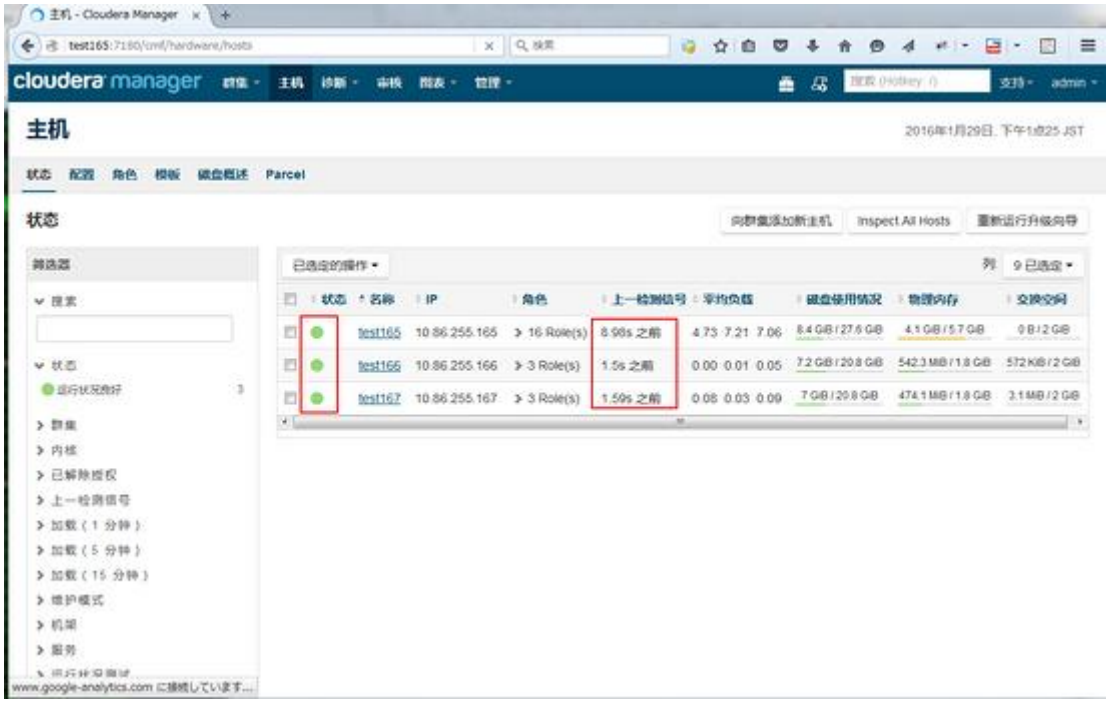


3. 确认，测试

确认集群状态正常，动作正常

1. 在集群页面确认，所有服务状态正常

2. 在主机页面确认，各节点的 Heartbeat 状态正常，并且时间小于 15 秒



3. 运行任务进行测试

登陆到集群中任意一台主机，执行下面任务(用 Hadoop 计算 PI 值，圆周率)

后面 2 个数字参数的含义： 10 指的是要运行 10 次 map 任务，100 指的是每个 map 任务，要投掷多少次，2 个参数的乘积就是总的投掷次数。

sudo -u hdfs hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 10 100

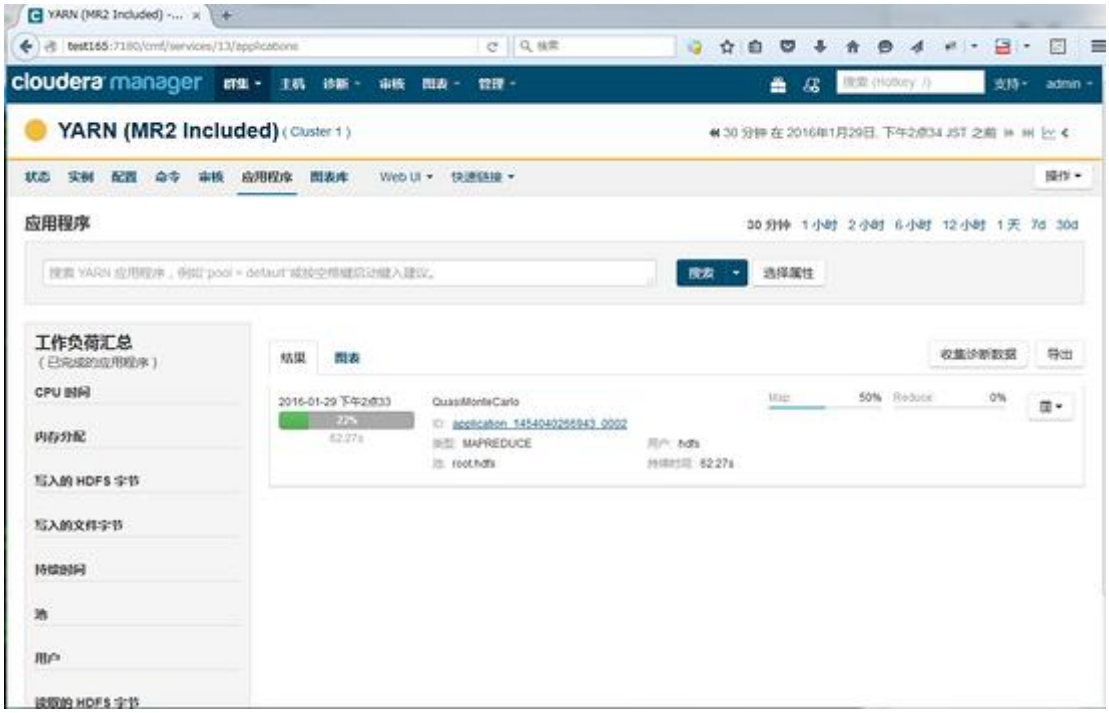
```
[root@test166 ~]# sudo -u hdfs hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar pi 10 100
Number of Maps  = 10
Samples per Map = 100
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Starting Job
16/01/29 13:31:11 INFO client.RMProxy: Connecting to ResourceManager at test165/10.86.255.165:8032
16/01/29 13:31:12 INFO input.FileInputFormat: Total input paths to process : 10
16/01/29 13:31:12 INFO mapreduce.JobSubmitter: number of splits:10
16/01/29 13:31:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1454040266943_0001
16/01/29 13:31:14 INFO impl.YarnClientImpl: Submitted application application_1454040266943_0001
16/01/29 13:31:14 INFO mapreduce.Job: The url to track the job: http://test165:8088/proxy/application_1454040266943_0001/
16/01/29 13:31:14 INFO mapreduce.Job: Running job: job_1454040266943_0001
16/01/29 13:31:27 INFO mapreduce.Job: Job job_1454040266943_0001 running in uber mode : false
16/01/29 13:31:27 INFO mapreduce.Job:  map 0% reduce 0%
16/01/29 13:31:39 INFO mapreduce.Job:  map 10% reduce 0%
16/01/29 13:31:48 INFO mapreduce.Job:  map 20% reduce 0%
16/01/29 13:31:57 INFO mapreduce.Job:  map 30% reduce 0%
16/01/29 13:32:06 INFO mapreduce.Job:  map 40% reduce 0%
16/01/29 13:32:14 INFO mapreduce.Job:  map 50% reduce 0%
16/01/29 13:32:22 INFO mapreduce.Job:  map 60% reduce 0%
16/01/29 13:32:31 INFO mapreduce.Job:  map 70% reduce 0%
16/01/29 13:32:39 INFO mapreduce.Job:  map 80% reduce 0%
16/01/29 13:32:47 INFO mapreduce.Job:  map 90% reduce 0%
16/01/29 13:32:55 INFO mapreduce.Job:  map 100% reduce 0%
16/01/29 13:33:05 INFO mapreduce.Job:  map 100% reduce 100%
16/01/29 13:33:05 INFO mapreduce.Job: Job job_1454040266943_0001 completed successfully
```

执行结果如下：

```
Job Finished in 113.953 seconds
Estimated value of Pi is 3.14800000000000000000
```

任务的执行情况可以在 YARN 页面上进行确认

群集 -> Cluster 1 -> YARN -> 应用程序

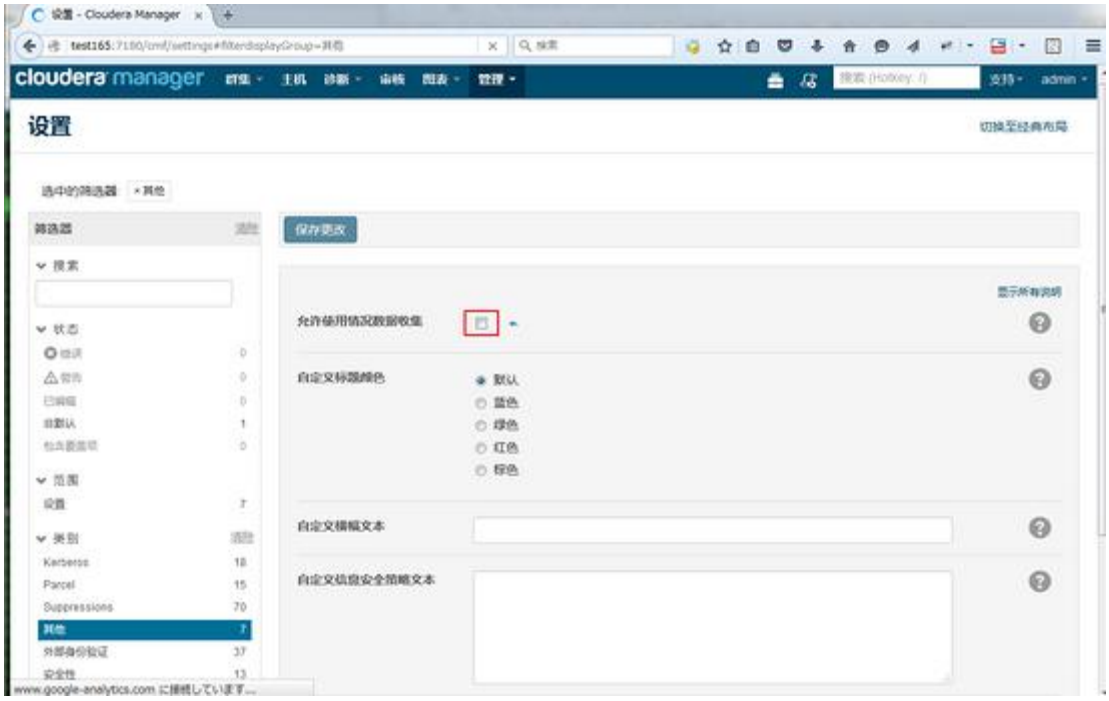


4. 其他

在 Cloudrea Manager 页面上，可以向集群中添加/删除主机，添加服务到集群等。

Cloudrea Manager 页面开启了 google-analytics，因为从国内访问很慢，可以关闭 google-analytics

管理 -> 设置 -> 其他 -> 允许使用情况数据收集 不选



5. 后记

工欲善其事必先利其器，管理 Hadoop 集群，Cloudrea 是个不错的选择。