

英特尔® Hadoop 发行版
版本 2.2
高可用性操作手册

目录

1. 简介.....	1
1.1 文档目的	1
1.2 使用到的包和软件	1
2. 配置要求.....	2
2.1 底层存储.....	2
2.2 用户 ID.....	2
3 集群配置.....	3
3.1 集群配置向导	3
3.2 集群拓扑配置	Error! Bookmark not defined.
3.3 更新许可证	Error! Bookmark not defined.
3.4 配置节点和格式化集群	14
3.5 集群组件的启动和停止	14
4 高可用性维护	16
4.1 DRBD 基本操作	16
4.1.1 DRBD 资源查看	16
4.1.2 查看 DRBD 状态.....	16
4.2 Pacemaker 基本操作	18
4.2.1 资源状态	19
4.2.2 启动和停止资源.....	19
4.2.3 对节点的操作.....	19
4.2.4 维护模式.....	19
4.2.5 手动修改 CRM 配置.....	20
4.3 故障处理后的主节点恢复	20
4.4 主从节点的同步恢复	21



1. 简介

1.1 文档目的

本文档用于指导英特尔® Hadoop 发行版高级用户如何配置、检查和维护英特尔® Hadoop 发行版的高可用性（HA）。

1.2 使用到的包和软件

- DRBD (Distributed Replicated Block Device), 是一款基于软件的、非共享的、主机之间块设备（硬盘、分区和逻辑卷等）中内容的复制存储方案。
- Pacemaker 是集群资源管理（cluster resource management, CRM）框架。它可以帮助您开始、停止、监视以及迁移资源。
- Corosync 是 Pacemaker 能够使用的集群通讯层。
- MySQL 是开源关系数据库管理系统（open source relation database management system, RDMBS）。



2. 配置要求

配置英特尔® Hadoop 发行版高可用性前，集群中的节点必须首先满足《英特尔® Hadoop 发行版新手指南》第二章中所介绍的系统要求，同时所有节点必须已经安装了符合要求的操作系统（详见《英特尔® Hadoop 发行版新手指南》第四章），管理节点必须已经正确安装了英特尔® Hadoop 发行版（见《英特尔® Hadoop 发行版新手指南》第五章）。在此基础上，还需要满足以下介绍的要求：

2.1 底层存储

您必须在主命名节点和备用命名节点上预留完全相同的存储区域，这将成为 DRBD 资源的底层存储。为了满足这一要求，您必须确保在主命名节点和备用命名节点上至少一个硬盘分区是完全等大的，并且不小于 64GB。

2.2 用户 ID

保证主命名节点和备用命名节点中的 hdfs/hadoop/mysql 这些用户 id 一致，否则 DRBD 分区的目录在主命名节点和备用命名节点切换之后会出问题。

3 集群配置

安装完英特尔® Hadoop 发行版后，管理员可以通过管理界面——Intel® Manager for Apache Hadoop 来完成高可用性配置。请通过浏览器访问英特尔® Hadoop 发行版管理界面地址 <https://管理节点地址:9443>，输入用户名和密码登陆。您可以在一开始配置集群时就选择高可用性组件，也可以为一个没有高可用性组件的、已配置好的集群配置高可用性。

如果您第一次配置集群，请接受英特尔® 软件最终用户许可协议，进入集群配置向导。

如果您要为一个已经配好的集群配置高可用性，请先停止所有正在运行的服务（请参见 [3.3 集群组件的启动和停止](#)）。然后点击界面中右上角的“配置向导”（如图 3.1 所示）进入集群配置向导。

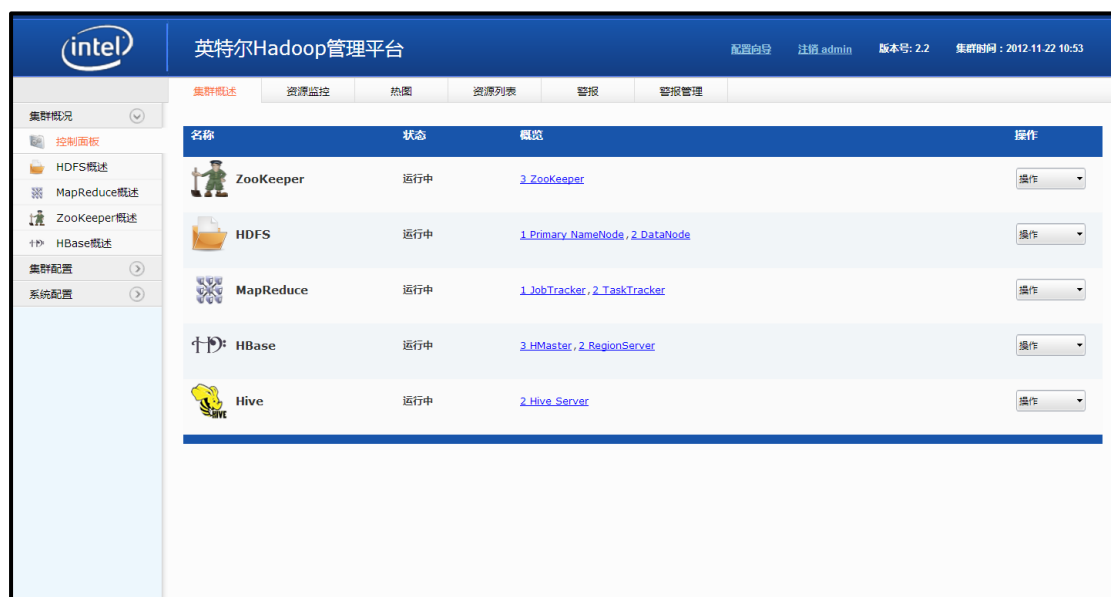


图 3.1 配置向导

3.1 集群配置向导

第一步，如下图所示在配置新的集群界面中输入集群名称和集群中所要安装的组件。请记住勾选“高可用性”。点击“下一步”继续。

3. 集群配置

第1步

配置新的集群

集群名称：

选择集群中将会使用的组件，包括HDFS，MapReduce，HBase，Hive，Sqoop，Pig和Flume；另外高可用性组件将会使用两台主备机器来保证集群的高可用性。

集群组件：

- HDFS：HDFS是一个分布式的文件系统。
- MapReduce：MapReduce是一种用于分布式系统的并行计算框架。
- ZooKeeper：ZooKeeper是一个针对大型分布式系统的可靠协调系统。
- HBase：HBases是基于HDFS的分布式的，可伸缩的，版本化的数据库系统。
- Hive：Hive是基于Hadoop的数据仓库工具。
- Sqoop：Sqoop是用于结构化数据存储和Hadoop之间的数据传输的工具。
- Pig：Pig是一个基于Hadoop的大规模数据分析平台。
- Flume：Flume是一个分布式的、可靠的、和高可用的海量日志聚合的系统。
- 高可用性：集群中将会有一台备份机器来保证高可用性。

图 3.2 配置新的集群

第二步，首先选择网络环境。然后您可以添加或删除节点。添加完成后点击“下一步”继续。

第2步

指定集群节点以及网络环境

网络环境：

节点名称	节点IP	状态
intelidh-01	192.168.1.71	已连通
intelidh-06	192.168.1.76	已连通
intelidh-07	192.168.1.77	已连通

图 3.3 “已连通”的状态

第三步，配置集群的机柜并点击“下一步”继续。

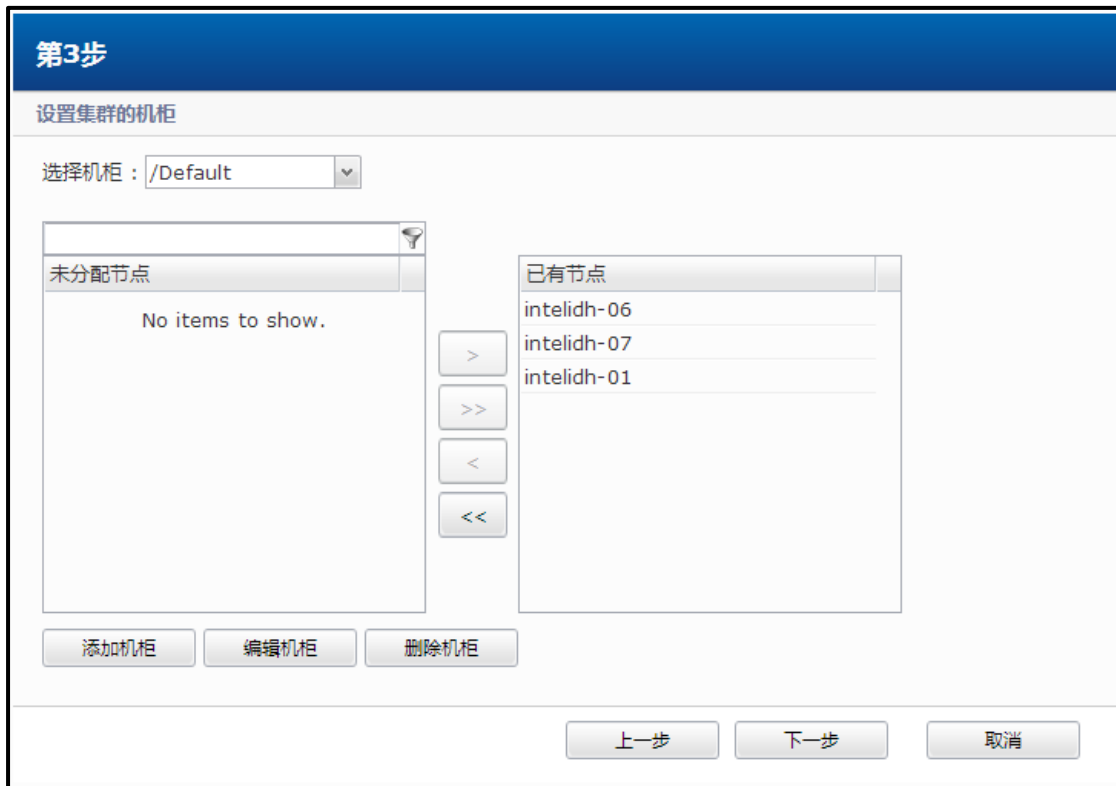


图 3.4 设置集群的机柜

第四步，选择安全策略并点击“下一步”继续。

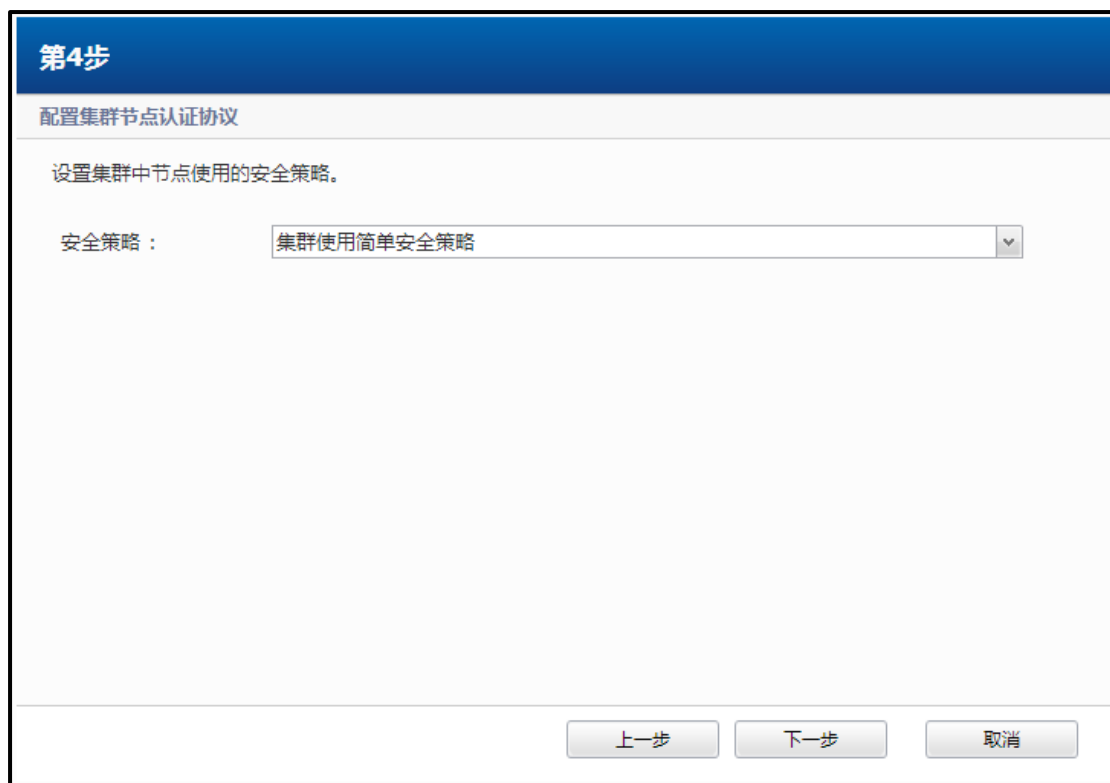


图 3.5 选择安全策略

第五步中，您可以配置集群的高可用性，如图所示。输入一个虚拟主机名和虚拟主机 IP。这个 IP 地址必须是不被集群中任何一个节点所使用的 IP 地址。然后选择一个服务器作为主命名节点和一个服务器作为备用命名节点，并选择选择他们的 DRBD 分区。这两个所选择的分区必须相同，并大于 64GB。虚拟主机名和虚拟 IP 是对外显示的主机名和 IP，其指向主命名节点或备用命名节点中活动的那个服务器。

3. 集群配置

第5步

HDFS高可用性配置

集群使用主备两台机器来保证集群的高可用性,当主机宕机时,备机会切换成主机接替原先主机的工作,从而保证集群正常运行不受影响。您需要指定主备两台机器所使用的虚拟主机名和IP地址;另外还需要指定这两台机器留作高可用性数据备份用的DRBD分区。

虚拟主机名 : Namenode的主机名,它是被主备两台机器共用的一个虚拟主机名。

虚拟IP地址 : 虚拟主机名对应的IP地址。**确保虚拟IP地址没有被局域网内其他机器使用。**

Primary NameNode : Primary NameNode DRBD分区 :

Standby NameNode : Standby NameNode DRBD分区 :

图 3.6 HDFS 高可用性配置

第六步,您可以配置 map/reduce 的高可用性,指定一台主备两台 Jobtracker 来确保主 Jobtracker 宕机时集群正常运行不受影响。

第6步

MapReduce高可用性配置

集群使用主备两台机器来保证MapReduce的高可用性,当主Jobtracker宕机时,备机会切换成主机接替原先主机的工作,从而保证集群正常运行不受影响。您需要指定主备两台Jobtracker所使用的虚拟主机名和IP地址。

虚拟主机名 : Jobtracker的主机名,它是被主备两台机器共用的一个虚拟主机名。

虚拟IP地址 : 虚拟主机名对应的IP地址。**确保虚拟IP地址没有被局域网内其他机器使用。**

Jobtracker :

Backup Jobtracker :

图 3.7 map/reduce 高可用性配置

第七步，如果有节点的状态为“未安装”，点击“安装未成功安装的节点”以将 Hadoop 软件安装到这些节点，并点击“是”确认选择。

如果所有节点的状态都为“成功”，点击“下一步”继续。

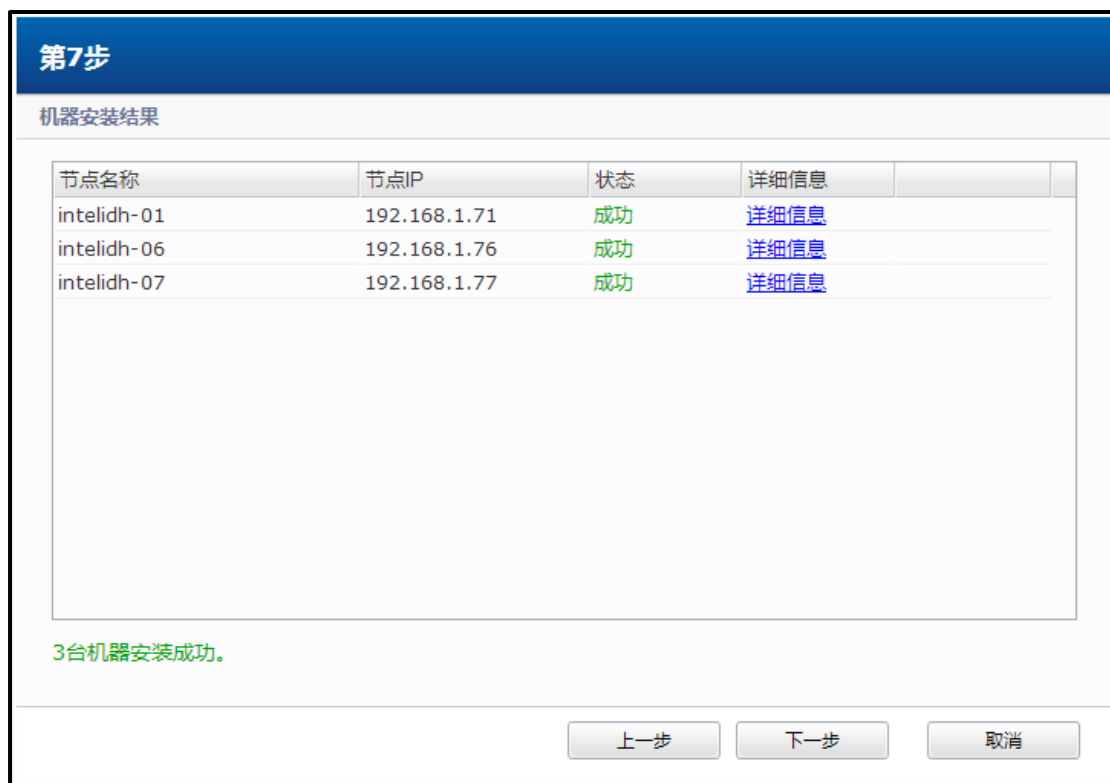


图 3.8 在集群的节点中安装 Hadoop 软件

第八步，开始进行集群拓扑配置。进入 HDFS 组件控制节点的配置界面。分别选择服务器选择作为主命名节点、从命名节点及备用命名节点。在配置高可用性时，您必须选择作为主命名节点和备用命名节点的服务器，而从命名节点是可选的。在这步当中，在集群配置向导第四步高可用性配置中您选择的主命名节点和备用命名节点将被默认作为 HDFS 中的主命名节点和备用命名节点。选择结束后，点击“下一步”继续。

第8步

HDFS组件控制节点的配置

Primary NameNode : (*)必填, 集群中必须包含一个Primary NameNode.

Secondary NameNode : 选填, Secondary Namenode可以备份Primary NameNode的元数据.

Standby NameNode : 选填, 如果集群使用高可用性则需要配置该项.

图 3.9 配置 HDFS 组件控制节点

进入 MapReduce 组件控制节点的配置界面。在这一步中，您在高可用性配置中选择的主命名节点将被默认作为 MapReduce 的任务分配器，备用命名节点将被默认作为 MapReduce 的备用任务分配器。点击“下一步”继续。

第9步

MapReduce组件控制节点的配置

JobTracker : (*)必填, 集群中必须包含一个JobTracker。

Backup JobTracker : 选填, 如果集群中使用了高可用性则需要配置该项。

图 3.10 MapReduce 组件控制节点配置界面

第九步, 进入 Zookeeper 组件控制节点配置界面。这里可以选择 ZooKeeper 节点, 建议使用奇数 Zookeeper 并且其数量至少为 3。选择结束后点击“下一步”继续。

3. 集群配置

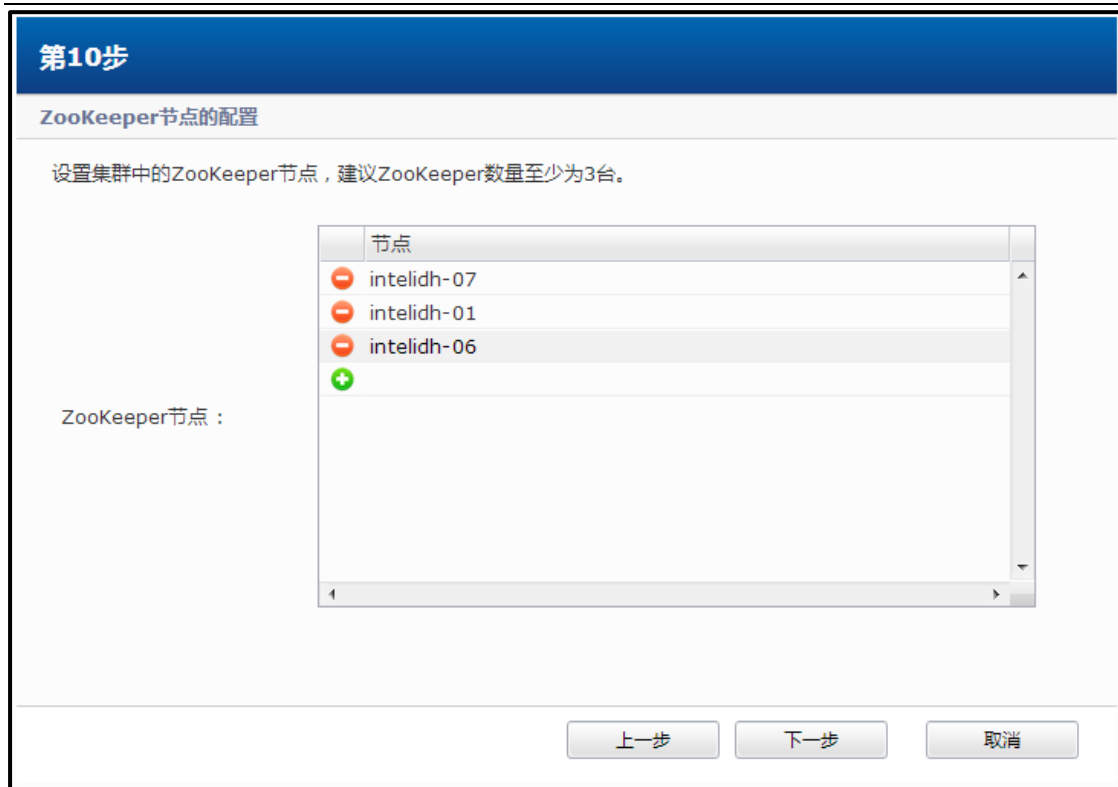


图 3.11 Zookeeper 组件控制节点配置界面

第十步，进入 HBase 组件控制节点配置界面。这里可以选择 Hbase 节点，建议与 Zookeeper 结点一致。选择结束后点击“下一步”继续。

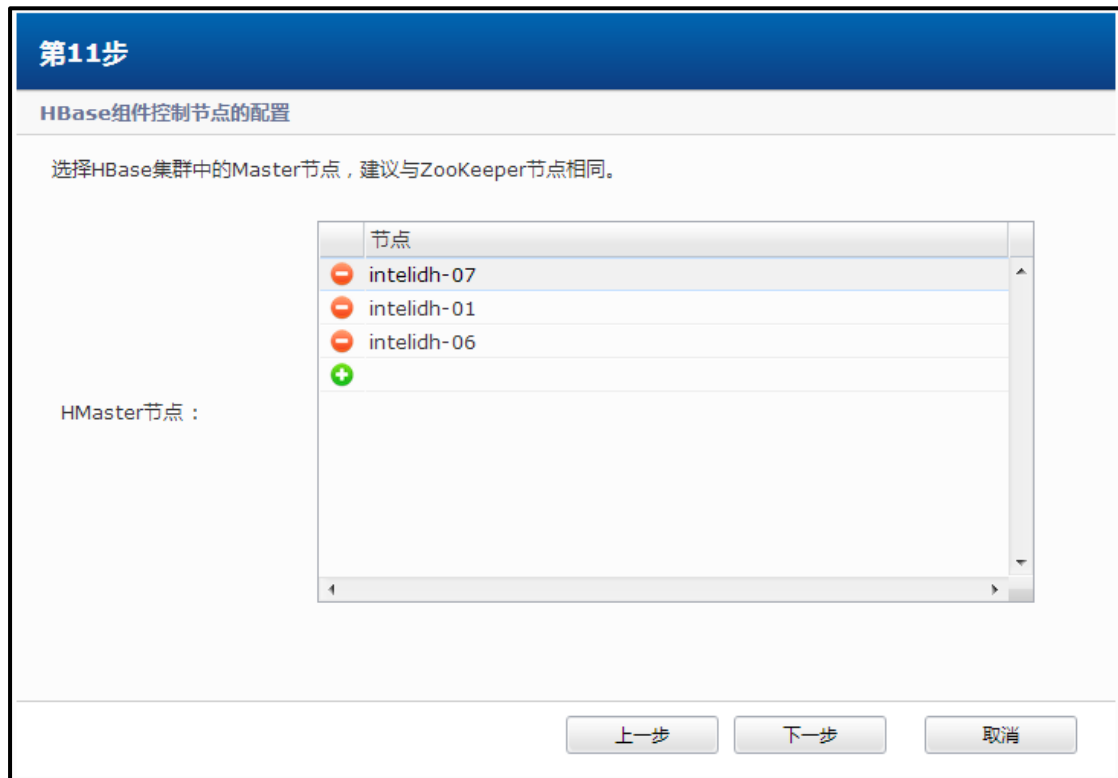


图 3.12 HBase 组件控制节点配置界面

第十一步，进入 Hive 组件控制节点配置界面。这里可以选择 Hive 服务所安装的服务器。选择结束后点击“下一步”继续。

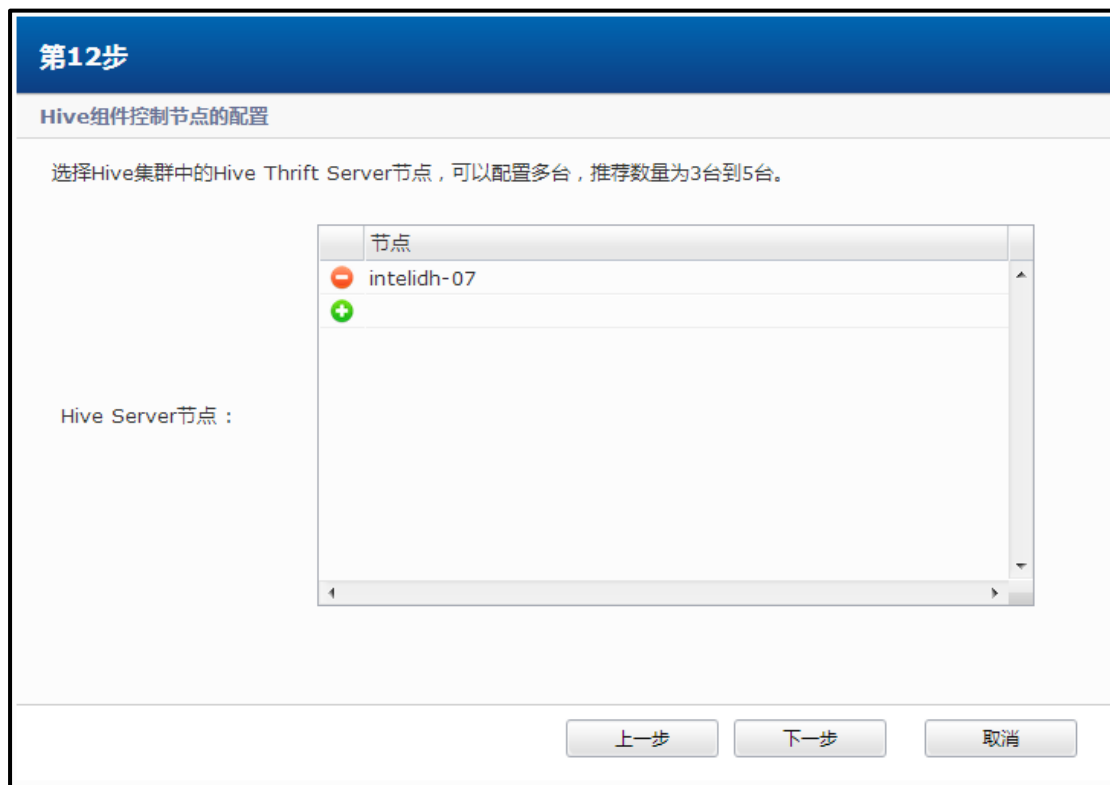


图 3.13 Hive 组件控制节点配置界面

第十二步，确认集群拓扑主要节点配置完成，点击“完成”关闭向导。

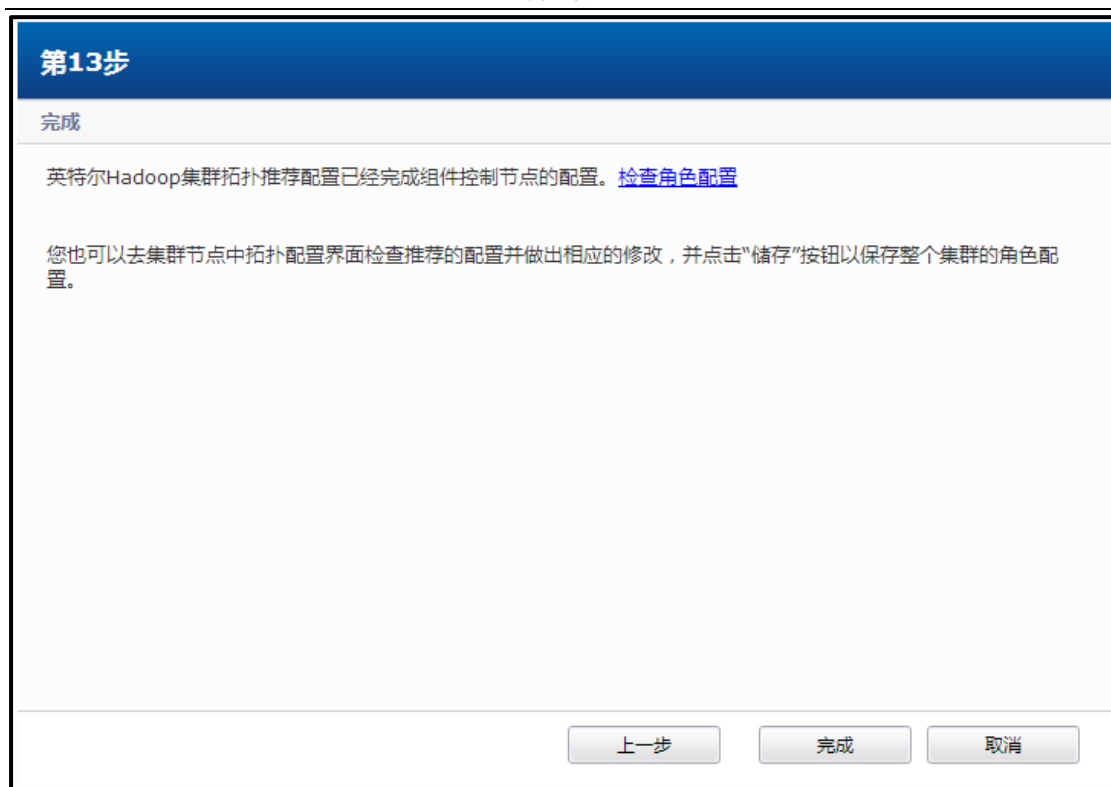


图 3.14 完成集群拓扑配置

如果是第一次安装，请点击确认进行“格式化集群并进行配置”，否则在节点配置中点击配置所有节点选项。



图 3.15 确认格式化集群配置

至此，集群安装配置已全部完成。

3.2 配置节点和格式化集群

如果您在安装配置后又修改了集群的配置，请点击如下图所圈出的“配置所有节点”。系统会同步在后台实施集群中各个服务器的安装配置工作。



图 3.20 配置所有节点

如果您是为一个已存在的、原来没有高可用性的集群配置高可用性，点击下图所示的“格式化集群”。



图 3.21 格式化集群

当所有服务器都启动完成后，点击启动过程界面右上角的关闭按钮关闭窗口。

3.3 集群组件的启动和停止

在运行中的组件分别为 Zookeeper，HDFS，MapReduce，HBase 和 Hive。除 HDFS 外，所有组件都只有两种状态“运行中”和“未运行”，在界面的最右侧有操作列表，可以通过点击按钮来对集群单一组件进行“启动/停止”操作。

单一组件的启动必须满足如下要求：

- HDFS，Zookeeper 不需要依赖另外组件；
- MapReduce 启动之前，需要确保 HDFS 处于运行状态下；
- HBase 启动之前，需要确保 HDFS，Zookeeper 处于运行状态下；
- Hive 启动之前，需要确保 HDFS，MapReduce 以及 HBase 处于运行状态下。

所以如需启动集群，需要严格按照启动顺序：Zookeeper，HDFS，MapReduce，HBase，Hive。

3. 集群配置

点击控制面板界面上相应控件旁的启动按钮，自上而下顺序启动每个服务。系统会显示每个服务启动的进度。启动完成后，系统会显示所有服务已经在运行中，证明系统安装成功。如下图所示。

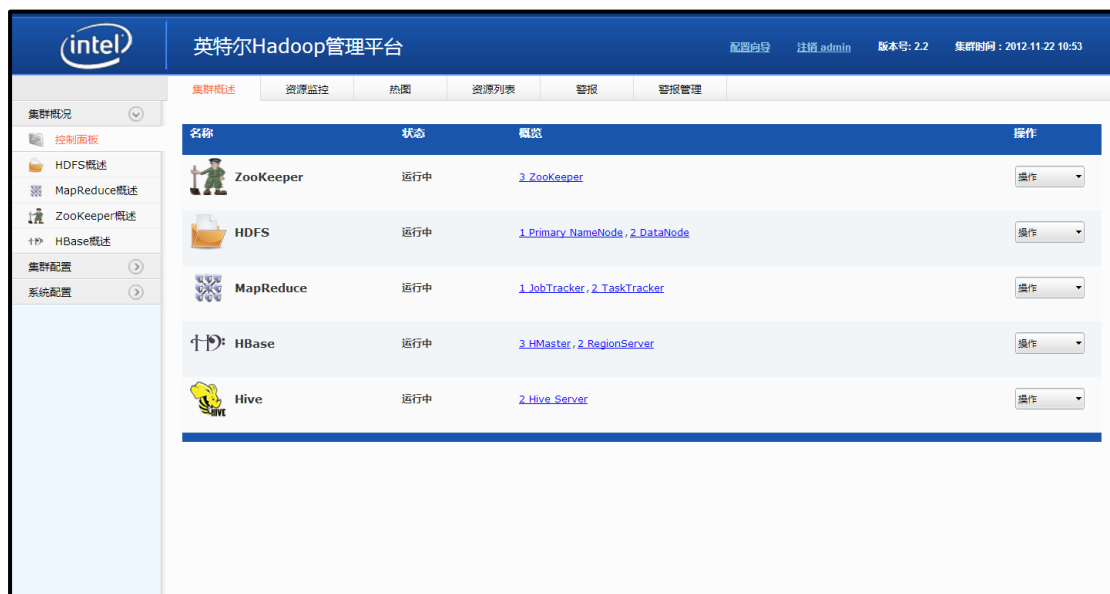


图 3.22 所有服务已经在运行

单一组件的启动必须满足如下要求：

- Zookeeper 启动之前不需要依赖另外组件；
 - HDFS 启动之前不需要依赖另外组件；
 - MapReduce 启动之前，需要确保 HDFS 处于运行状态下；
 - HBase 启动之前，需要确保 Zookeeper，HDFS 处于运行状态下；
 - Hive 启动之前，需要确保 HDFS，MapReduce 以及 HBase 处于运行状态下。
- 所以如需启动集群，需要严格按照启动顺序：Zookeeper，HDFS，MapReduce，HBase，Hive。点击控制面板界面上相应控件旁的启动按钮，自上而下顺序启动每个服务。

4 高可用性维护

4.1 DRBD 基本操作

4.1.1 DRBD 资源查看

在 DRBD 中,资源定义为一个特定复制数据集各方面的统称,包括资源名称、数量和 BRDB 设备。

DRBD 资源的配置文件存在于/etc/drbd.d 目录。配置文件的名字与资源的名字是一样的。举例来说,您可以用一下命令查看定义为 r0 资源的配置内容:

```
vi /etc/drbd.d/r0.res
```

其配置内容的形式如下:

```
resource r0
{
    device /dev/drbd0;

    on xtt-portal
    {
        disk          /dev/sdb1;
        address       10.239.47.81:7789;
        meta-disk     internal;
    }

    on xmlqa-clv9
    {
        disk          /dev/sdb1;
        address       10.239.47.35:7789;
        meta-disk     internal;
    }
}
```

这个例子对 DRBD 进行的如下配置:

- 作为高可用性机对包括两个节点, xtt-portal 和 xml-qa-clv9。
- 命名为 r0 的资源使用/dev/sdb1 作为底层存储, 并且配置了内部数据元。
- 资源使用 TCP 端口 7789 作为其网络连接, 并分别和 10.239.47.81 和 10.239.47.35 这两个 IP 地址绑定。

4.1.2 查看 DRBD 状态

查看节点的 DRBD 信息可以在相应节点执行命令:



4. 高可用性维护

```
cat /proc/drbd
```

表 4.1、表 4.2、表 4.3 和表 4.4 能帮助你读懂 DRBD 资源状态的含义。

缩写	含义	命令	状态
cs	连接状态 connection state	drbdadm cstate <资源>	(详见表 4.2)
ro	角色 roles	drbdadm role <资源>	本机资源角色/ 另一台 子资源角色 (详见表 4.3)
ds	磁盘状态 disk state	drbdadm dstate <资源>	本机磁盘状态/ 另一台 子资源磁盘状态 (详见表 4.4)
p	复制协议 replication protocol		A, B 或 C

表 4.1 DRBD 信息中缩写及其含义

状态	含义
StandAlone	没有可用的网络配置。资源并没有被连接或已经被管理员断开,或因为验证失败或脑裂而断开连接。
WFConnection	这一个节点正在等待网络中另一个节点
WFReportParams	TCP 连接已经被建立, 这个节点正在等待另一个节点发送的第一个网络数据包
Connected	DRBD 连接已经建立, 这是正常工作的状态。

表 4.2 主要的连接状态

角色	含义
Primary	该资源当前的角色为主, 可以被读写。除非双主模式被启用, 否则在同一时刻, 两个节点中只能有一个节点被赋予该角色。
Secondary	该资源当前的角色为从, 可以正常接收另一节点发送过来的更新(除非其处于未连接的状态), 但不能被读写。机对中一个或两个节点都可以同时作为该角色。
Unknown	该资源当前的角色未知。本机资源不会处于该角色。只有另一节点处于未连接状态时, 它的角色才有可能为Unknown。

表 4.3 主要的角色及其含义

磁盘状态	含义
Diskless	没有本地的块设备被分配到 DRBD 驱动。这意味着资源没有被



4. 高可用性维护

	附加到其备份设备, 这可能是因为其被 <code>drbdadm detach</code> 这一命令分离了, 或因为底层 I/O 错误被自动分离。
Inconsistent	数据不一致。当新建一个资源时, 两个节点会马上都变为这一状态, 并在同步完成前保持该状态。另一种情况, 在同步时, 同步目标的那个节点也会处于该状态。
Outdated	资源数据是一致的, 但已经过期。
DUnknown	当网络连接不可用是, 另一磁盘将处于该状态。
Consistent	没有连接时数据一致。当连接后, 该状态将会变为 UpToDate 或 Outdated。
UpToDate	数据一致并已更新。这是正常的状态。

表 4.4 主要的磁盘状态及其含义

亦可执行如下命令以得到更简洁的 DRBD 信息:

```
drbd-overview
```

4.1.3 常用命令

表 4.2 列出了 DRBD 常用命令及其作用。

命令	作用
<code>drbdadm up <资源></code>	在主机上启用 DRBD 资源
<code>drbdadm down <资源></code>	禁用 DRBD 资源
<code>drbdadm primary <资源></code>	把一个 DRBD 资源转为 primary 模式
<code>drbdadm secondary <资源></code>	把一个 DRBD 资源转为 secondary 模式
<code>drbdadm attach <资源></code>	将一个资源附加到其备份设备
<code>drbdadm detach <资源></code>	把一个资源从其备份设备中分离出来
<code>drbdadm connect <资源></code>	启用 DRBD 资源的网络连接
<code>drbdadm disconnect <资源></code>	禁用 DRBD 资源的网络连接
<code>drbdadm pause-sync <资源></code>	停止正在运行的再同步过程
<code>drbdadm resume-sync <资源></code>	恢复再同步过程

表 4.2 DRBD 的常用命令及作用

4.2 Pacemaker 基本操作

本节简单介绍了与高可用性有关的几个常用的基本操作。更多操作的详细信息您可以通过

```
crm help <命令>
```

来获得。



4.2.1 资源状态

资源包括: fs_hadoop, ms_drbd_hadoop (drbd_hadoop), namenode, ip_hadoop, hive, mysqld, hive_metastore 和 jobtracker。

如果您想要得到所有资源的当前状态的概览, 您可以使用如下命令:

```
crm resource show
```

如果您需要更详细的信息, 请使用如下命令:

```
crm_mon -l -rf
```

这一命令将列出包括节点、错误统计及非活动资源等信息。如果把“-l”省略, 每有事件更新, 新的信息将被列出。

4.2.2 启动和停止资源

启动资源的命令如下:

```
crm resource start <资源名称>
```

停止资源的命令如下:

```
crm resource stop <资源名称>
```

4.2.3 对节点的操作

如果您想要将节点成为 Standby 模式, 请使用如下命令:

```
crm node standby <节点名称>
```

要使节点进入 Online 状态, 并允许 Pacemaker 在节点上启用相关资源, 请使用如下命令:

```
crm node online <节点名称>
```

如果您是在需要更面状态的节点上运行该命令, 节点名称可以省略。

4.2.4 维护模式

如果您想要维护您的集群并希望 Pacemaker 暂时不干预集群的工作, 您可以使用以下命令启用维护模式:

```
crm configure property maintenance-mode="false"
```

关闭维护模式的命令如下:



4. 高可用性维护

```
crm configure property maintenance-mode="false"
```

4.2.5 手动修改 CRM 配置

如果您想要手动修改 CRM 配置，您可以使用以下命令进去配置界面进行修改：

```
crm configure  
edit
```

4.3 故障处理后的主节点恢复

处理故障后，主节点可能发生变化，如果希望恢复原来的 master 节点，需要手动修改：

1. 在需要成为 master 的节点上执行命令：

```
service drbd start  
service corosync start  
service pacemaker start  
crm node online
```

2. 找到现在的 master 节点，在 master 节点上输入命令：

```
crm node standby
```

3. 等待主节点发生切换结束，可以通过 `crm status` 命令查看切换是否完成，并且保证基本服务已经成功启动，包括：

- ① DRBD 资源，查看命令如下：

```
crm resource status ms_drbd_hadoop
```

- ② Master 节点上的 Drbd 分区是否成功挂载到 `/hadoop/drbd` 目录，查看命令如下：

```
crm resource status fs_hadoop
```

- ③ 虚拟 IP 是否成功绑定到 Master 节点，查看命令如下：

```
crm resource status ip_hadoop
```

4. 切换完成之后的从节点上执行命令：



4. 高可用性维护

```
crm node online
```

4.4 主从节点的同步恢复

执行如下命令：

```
vi /var/log/messages
```

如果在中出现类似信息：

```
Split-Brain detected, dropping connection!
```

则发生脑裂故障，Primary 节点和 Standby 节点两个节点间数据不再同步。

检查，在 Standby 节点机器上命令行输入命令：

```
service drbd status
```

这是您必须删除不一致数据。在 Standby 节点机器上命令行输入：

```
drbdadm disconnect r0  
drbdadm secondary r0  
drbdadm -- --discard-my-data connect r0
```

在 Primary 节点上重连接资源，在 Primary 节点机器命令行输入：

```
drbdadm connect r0
```

在 Standby 节点上再次启动 drbd 服务：

```
service drbd start
```

再次查看 drbd 服务，查看状态是否正常：

```
service drbd status
```