Cloudera 产品高可用性配置说明

包含 HDFS HA，以及配置 CDH 其他组件如 Hive Metastore、Hue、Impala 使用 HDFS HA

cloudera
Ask Bigger Questions

Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304 | 1-888-789-1488 or 1-650-362-0488 | cloudera.com

©2013 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera, Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

版本	作者	日期	描述
1.0	李大超 （dachao.li@）	2014-11-5	适用于 CDH5.0 及以上版本

cloudera
Ask Bigger Questions

Cloudera, Inc. 220 Portage Avenue, Palo Alto, CA 94306 USA | 1-888-789-1488 or 1-650-362-0488 | cloudera.com

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

目录

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

1. 简介

Apache Hadoop*集群中通常有多个用户长时间地运行多个作业。这些作业生成的数据分析具有商业上的重要性，可帮助公司节省大笔开支或产生收入。因此，集群的高可用性至关重要，几分钟、几小时或几天的宕机可能花费大量的金钱。

系统管理员面对的问题主要是 Primary NameNode 的单点故障。如果其中一个服务失败，则在问题解决前集群功能将不可用。而且，这些故障可能需要花费大量的时间和人力去解决，这将导致长时间宕机，这对公司业务尤其是关键业务来说是不可接受的。

要解决这些问题，Cloudera 产品支持 HDFS 的高可用性（High Availability）功能。HDFS 高可用性是 Apache Hadoop*实施的一个开源解决方案。

本文将介绍如何为 Cloudera 产品设置高可用性。

2. 了解高可用性架构

高可用性功能支持 Primary NameNode 的 active-standby 配置。这表示 Primary NameNode 在另一个节点上有一个完全冗余的对象，它只有当 Primary NameNode 发生故障时会被激活。Primary NameNode 的冗余对象被称为 Standby NameNode。

Primary NameNode 负责集群中的 HDFS 操作，比如从 HDFS 读取文件并写入文件到 HDFS。Standby Namenode 的作用是维护 HDFS 集群的状态，以便提供热备份。热备份是指如果 Primary NameNode 发生故障时，能立即切换到 Standby NameNode 而不会产生服务中断的情况。

2.1 高可用性设计概述

要成为热备份，Standby Namenode 必须对以下数据有连续的、即时的读取权限：

HDFS 命名空间的更改，比如重命名、删除或创建文件。

Primary NameNode 存储编辑日志到一个名为 Shared Edits 的特定目录下。Standby NameNode 对此目录的文件有读的权限，因此可根据存储在编辑日志中的数据来更新 HDFS 结构。这意味着对 Primary NameNode 命名空间做出的任何更改都将被复制到 Standby NameNode 的命名空间。

©2013 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera, Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

DataNode 已被配置为可同时发送块位置信息到 Primary NameNode 和 Standby NameNode。

2.2 处理 HDFS 命名空间的更改

当客户端在 HDFS 上执行写操作时，这一事件将首先被记录在预写式日志，或编辑日志。一旦编辑日志更改成功，Primary NameNode 的文件系统结构的内存中信息将被更新。Standby NameNode 文件系统结构必须和 Primary NameNode 的文件系统结构完全相同。这意味着 Standby Namenode 必须对 Primary NameNode 的编辑日志有读的权限。

高可用性设计使用 *shared edits* 目录来达到这一要求。此目录是 Primary NameNode 存储和更新编辑日志文件的目录，也是 Standby NameNode 读取编辑日志的目录。Standby NameNode 使用编辑日志中的信息来更新 HDFS 命名空间的内存中信息。此外，Standby NameNode 不可改动编辑日志，只能读取。通过这些操作，Standby NameNode 确保了 HDFS 文件结构会一直和 Primary NameNode 保持一致。

如果发生故障切换，Standby NameNode 将确认在激活前它已读取所有编辑日志中的信息并更新了命名空间。因此，Standby NameNode 在成为 Primary NameNode 之前，它的命名空间将保持和 Primary NameNode 同步。

2.3 访问 Shared Edits 目录的机制

Shared Edits 目录必须可被二个 NameNode 访问，且都对此目录的文件有读 / 写权限。此外，NameNode 必须能不间断地读 / 写此目录，且二个 NameNode 都能一直访问同样的数据。高可用性 HDFS 支持以下二种授权 NameNode 访问 Shared Edits 目录的方式：

- 网络文件共享（NFS）
- Quorum Journal Manager

2.3.1 为共享存储使用 NFS

Shared Edits 目录可被放在服务器的某个目录，即 NFS 挂载 NameNode 的服务器。你可能只有一个 Shared Edits 目录，因此 NFS 将会成为 HDFS 单点故障所在点。如果 NFS 出现问题，则 HDFS 客户端将不能写入数据到 HDFS。因此，存储 Shared Edits 目录的服务器应被配置为高可用、高质量的专用 NAS 设备。

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

2.3.2 为共享存储使用 Quorum Journal Manager

在多数组织中，将 Shared Edits 目录存放在挂载的 NFS 上是可被接受的方式，这通常也符合公司现存的整体架构。但是，对有些组织来说，NFS 挂载选项可能造成以下问题。

- 成本和客户硬件问题 — 由于 NFS 挂载可能是 HDFS 单点故障（SPOF）所在点，强烈建议用户将 Shared Edits 目录放在 NAS 设备上。NAS 设备通常很昂贵，而且要求特别的设备用于维护和操作。

- 操作、实施和管理 —除了部署 HDFS 外，NFS 挂载还要求额外的配置、监控和维护。这将增加高可用性设置的复杂性，并可能导致 NFS 配置错误。如果 NFS 配置错误，这将导致 HDFS 不能工作。最后，存放 Shared Edits 目录的服务器或 NAS 设备将增加机构对外部设备的依赖性，这一点对管理和维护 NameNode 影响重大。

- NFS 客户端有很多缺陷—在 Linux 操作系统中，NFS 客户端尚有很多缺陷，且难以配置。而且，每个客户端的具体实施可能不尽相同，也就是说，预测客户端的行为将很困难，而且每个客户端的行为都不相同。因此，将 NFS 客户端挂载在 Shared Edits 目录很容易导致错误，这将使得 NameNode 不能读或写该目录。

要解决这一问题，Apache Hadoop*社区用 QuorumJournalManager （QJM）作为 NFS 选项外的另一个选择。QuorumJournalManager 是一个分布式应用程序，它将相同的 HDFS 编辑日志存储在集群的多个节点上。

2.4 QuorumJournalManager 的功能

QuorumJournalManager 设计具有 NFS 不具备的以下优点：

- 没有单点故障 — 如果 QuorumJournalManager 集群中的一个或多个节点宕机，则 HDFS 编辑日志的备份将在集群中的另一个节点上变为可用。QJM 集群中允许失败且失败后可继续提供服务的节点数量根据以下算法计算： (N - 1) / 2

  例如，如果集群中有 5 个 JournalNode，则允许失败的节点数量是 2，且失败后可继续提供服务。例如，如果集群中有 3 个 JournalNode，则允许失败且的节点数量是 1，且失败后可继续提供服务。
- 无需特殊硬件或外部依赖 — 不像 NFS 通常要求有 NAS 设备，这一选项对硬件或服务无特殊要求。Quorum Journal Manager 可部署在部署了饔的同一标准硬件上。实际操作中，推荐将 Quorum Journal Manager 安装在已安装饔软件的同一节点上。

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

- 无需硬件隔离 — 在故障恢复中，隔离之前运行的 NameNode 是通过软件完成的。
- 用户可定义故障恢复节点数 — QuorumJournalManager 可定义集群中的故障恢复节点数。每个节点上都有和其它节点上一样的数据备份。因此，你可指定集群的大小，且可指定可用节点数目以提供服务。
- 网络延迟将影响所有节点 — QuorumJournalManager 集群中节点的编辑日志数据的复制不会影响对编辑日志数据的读 / 写速度。而且，增加要复制编辑日志数据的节点数会在 NameNode 读取数据时造成网络延迟。

2.5 QuorumJournalManager 设计概述

应用程序包含由数个节点组成的集群，HDFS 编辑日志存储在每个节点上。在 QuorumJournalManager 集群中，节点被称为 JournalNode。QuorumJournalManager 使用分布式协议来确保每个节点上的数据在任何时候都和其他节点同步，且只有激活的 NameNode 可编辑 HDFS 编辑日志。

NameNode 是消耗 QuorumJournalManager 资源的客户端。每个 NameNode 上都运行 QuorumJournalManager 服务。Primary NameNode 使用 QJM 服务来对一组 JournalNode（JN）进行写入编辑操作。QJM 在多数节点返回成功消息后，将认为写入编辑已成功。一旦集群确认编辑完成，编辑日志的某一部分可能只能从 QuorumJournalManager 集群读取。因此，Standby NameNode 可使用 QJM 来读取集群中任何节点上的编辑日志复制数据，并确保该复制数据和集群中其他的复制数据完全一致。

2.6 发送块位置信息到 NameNode

要让 Standby NameNode 成为热备份，必须保证它有最新的所有 DataNode 上所有文件的块位置信息。在 HDFS 中，DataNode 负责向 NameNode 周期性报告存储在其中的快信息。NameNode 不负责也不读取编辑日志的块位置信息。因此，高可用性设计中将通过改变 DataNode 报告块信息的方式来完成这一点。

在高可用性配置中，Primary NameNode 和 Standby NameNode 的网络地址在每个节点的 hdfs-site.xml 文件中被定义。通过配置文件中定义的网络地址，DataNode 发出数据块报告、块位置更新信息和心跳到这二个 NameNode。但是，DataNode 仅执行 Primary NameNode 发出的数据块有关命令。

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

2.7 和 NameNode 的客户端通信

HDFS 高可用性不支持 NameNode 的 active-active 配置模式。这表示在任何时候，只能有一个 NameNode 用于管理 HDFS 命名空间和处理 HDFS 客户端请求。为确保 HDFS 客户端仅和 active 状态的 NameNode 通讯，客户端将会：

- HDFS 客户端有存储在配置文件中的每个 NameNode 的网络地址。
- 客户端将尝试和配置文件中的第一个地址通信。
- 如果客户端返回 Standby NameNode 的消息，则服务将返回消息表示这是一个 Standby NameNode，客户端需要尝试和另一个 NameNode 通信。
- 如果客户端收到上一步所述消息，则它将放弃和 Standby NameNode 的通信，然后和另一个 NameNode 建立通信。

以上过程将重复直至找到 active 状态的 NameNode。如果客户端不能找到 active 状态的 NameNode，则通信将失败并终止。

2.8 NameNode 故障恢复

故障恢复从 Primary NameNode 切换到 Standby NameNode 时，可以是人工或自动操作。以下内容描述了每种故障恢复模式。

2.8.1 人工故障恢复

在高可用性 HDFS 集群中，管理员可使用以下新命令：hdfs haadmin。这一命令有一个名为 -failover 的子命令。failover 子命令有二个额外的参数。第一个参数是当前 active 状态的 NameNode 的逻辑名，第二个参数是当前 Standby NameNode 的逻辑名。执行命令后，故障恢复尝试从 Primary NameNode 切换到 Standby NameNode。

2.8.2 自动故障恢复

热备份的一个重要特征是当 active 状态的节点不能提供服务时进行实时故障恢复。要达到这一要求，高可用性 HDFS 功能支持自动故障恢复。自动故障恢复是指当服务探测到 active 状态的节点不能提供服务时，通过程序或服务立即触发故障恢复。

如果启用了自动故障恢复，高可用性配置中将增加二个组件。这些组件负责处理自动故障恢复。

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

- ZooKeeper quorum
- ZKFailoverController （ZKFC）

Apache ZooKeeper (ZK) *是一个高可用性协调服务，用于维护分布式应用程序的状态、进程和配置数据。ZooKeeper quorum 是 ZooKeeper 集群中的一组节点，用来存储单个应用程序的数据。在高可用性自动故障恢复中，ZooKeeper quorum 提供以下功能：

- 探测 NameNode 的失败 — 每个 NameNode 的状态在 ZK quorum 都作为持续进程进行维护。如果有 NameNode 进入无法提供服务的状态，则 ZK 进程将失效，并通知另一 NameNode 此状态。如果无法提供服务的 NameNode 之前为 active 状态，则其他 NameNode 将触发故障恢复，使自己成为 active 状态的节点。
- 为 NameNode 的推选和状态设置锁 — 要表示哪个 NameNode 是 active 状态的，当前 active 状态的 NameNode 会在 ZK quorum 中被锁定。锁定表示了二个 NameNode 中哪个 NameNode 是 active 状态的。NameNode 只有在获得此 ZK 锁之后才会将状态改变为非 active 或 active。

如果 active 状态的 NameNode 进入非服务状态（OOS），这将导致另一个 NameNode 触发故障恢复。如果 active 状态的 NameNode 失败，则 Standby NameNode 将获得 ZK 锁然后成为 active 状态的 NameNode。因此，如果 active 状态的 NameNode 进入非服务状态（OOS），Standby NameNode 开始故障恢复，且 active 状态的 NameNode 在故障恢复过程中重新回到服务状态，在此情形下由于有 ZK 锁，Standby NameNode 仍将成为 active 状态的 NameNode。

ZKFC 是指 ZooKeeper 客户端，用于运行二个 NameNode。ZK 客户端提供以下功能：

- 状态监控 — 定期发送心跳到 NameNode，以检查 NameNode 是否运行状态良好。如果 NameNode 及时对发送的心跳返回状态良好的信息，则 ZKFC 将认为 NameNode 可用且状态良好。如果 NameNode 没有在一定的时间内返回信息或返回的状态为非良好状态，则 ZKFC 将认为 NameNode 状态错误。
- ZooKeeper 进程和锁定管理 — 一旦 ZKFC 认为 NameNode 状态良好，则 ZKFC 将在 ZK quorum 中对此维护一个持续进程。此进程表示 NameNode 状态良好且可用。如果 ZKFC 判定 NameNode 状态错误，则 ZKFC 将在 ZK quorum 中终止这一进程。如果进程终止且 NameNode 当前为 active 状态，则用于 active 状态的 ZK 锁将被释放。
- 锁必须永远在其中一个 NameNode 上。

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

- Active 状态的 NameNode 推选 —如果 ZKFC 判定 NameNode 运行状态良好，且其他 NameNode 没有 ZK 锁，则 ZKFC 将为 NameNode 尝试获取锁。如果 ZKFC 成功获得锁，则 ZKFC 必须触发一个故障恢复以使得 NameNode 成为 active 状态的 NameNode。

2.9 通过隔离防止脑裂现象

高可用性 HDFS 不支持双机热配置，也就是说，同一时间内只有一个 NameNode 可成为 active 状态。如果二个 NameNode 同时成为 active 状态，则每个节点上的 HDFS 命名空间将与对方快速分离。如果 HDFS 结构分离，则很有可能 HDFS 上的数据会丢失或损坏。而且，HDFS 客户端可在二个 NameNode 上执行同一操作，但从一个 NameNode 返回的结果可能和另一个 NameNode 返回的结果明显不同。

脑裂现象的产生是由于二个节点都认为在 active-standby 集群中自己是 active 状态的节点。以 HDFS 为例，这意味这二个 NameNode 都认为自己是 active 状态的节点，其他节点是 Standby NameNode。要确保高可用性 HDFS 中没有数据丢失、损坏或不一致，防止脑裂现象非常重要。

在集群中，隔离进程用于将工作不正常的服务孤立起来，以防止该服务访问共享资源。在高可用性 HDFS 中，隔离进程发生在 NameNode 推选之后，新的 active 状态的 NameNode 会验证之前为 active 状态的 NameNode 将不能编辑 HDFS 命名空间。

2.9.1 Fencing of the Shared Edit Directory on a NFS

为防止脑裂现象，高可用性 HDFS 在故障恢复成功前将执行以下隔离进程：

- 尝试验证之前为 active 状态的 NameNode 已不再是 active 状态。如果成功验证之前为 active 状态的 NameNode 已不再是 active 状态，则其他 NameNode 将成为 active 状态。
- 如果不能验证之前为 active 状态的 NameNode 是否还是 active 状态，则启用隔离机制。隔离操作将尝试找到并终止之前为 active 状态的 NameNode 服务，从而防止之前为 active 状态的 NameNode 访问 Shared Edits 目录。
- 如果隔离机制成功，则其他 NameNode 将成为 active 状态。如果隔离机制不成功，则故障恢复不会发生。

2.9.2 Fencing of the Shared Edit Directory in QuorumJournalManager

隔离是建立在 QJM 上的，它不需要额外的或特别的硬件和软件。在 QJM 中，NameNode 被认为是写入者，即对编辑日志进行更改的节点。要确保只有 active 状态的 NameNode 被允许更

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

改编辑日志，当 NameNode 成为 active 时，QJM 将会被分配一个纪元号（epoch number）。而且，写入者将不可写入编辑日志，直至确认之前所有的写入者都已停止写入该编辑日志文件。

纪元号是指带有以下属性的整数：

- 纪元号同时存储在 QJM 和所有的 JournalNode 上。
- 对于二个 NameNode 来说，纪元号具有唯一性。二个 NameNode 永远不可能有同一个纪元号。
- 要生成纪元号，NameNode 的 QJM 从所有的 JournalNode 上获取纪元号。QJM 将找出最高的纪元号，加上 1，结果即为新的纪元号。
- 纪元号定义了写入者的顺序，这样 QJM 和 JN 可决定写入者比之前的写入者新还是旧。如果 NameNode 的纪元号比其他 NameNode 的纪元号高，则纪元号较高的 NameNode 相对于另一个 NameNode，被认为是新的写入者。

QJM 通过以下方式使用纪元号：

- 在 NameNode 被允许更改编辑日志前，QJM 必须已被成功分配纪元号。
- 当 NameNode 成为 active 状态时，纪元号会生成并分配给该 NameNode。HDFS 命名空间第一次被格式化后，第一个 active 状态的 NameNode 将被分配纪元号 1。任何故障恢复都将导致纪元号的增加。
- 在纪元号被成功分配给 QJM 前，QJM 必须发送纪元号到集群中的所有 JournalNode 上。多数 JournalNode 必须返回一个消息，标明纪元号已成功收到，否则 QJM 将不能使用该纪元号。
- 如果 QJM 回应纪元号的消息请求，它将存储这一纪元号以便日后参考。无论何时 QJM 发送写请求到 JN，纪元号都包含在请求中。
- 当 JN 收到 QJM 发送的写请求时，JN 会将消息请求中的纪元号和存储在本地的纪元号进行对比。如果请求中的纪元号低于 JN 的纪元号，JN 将拒绝写请求。如果请求中的纪元号高于 JN 的纪元号，JN 将更新纪元号以使之和请求中的纪元号匹配并允许写入。

根据 QJM 使用纪元号的方式，脑裂现象将通过以下方式处理：

- 如果 NameNode 能成功写入到多数 JournalNode，NameNode 只可写入到编辑日志。因此，多数 JournalNode 必须接受 NameNode 的纪元号作为新的纪元号。
- 当 NameNode 成为 active 状态时，它的纪元号总是比其他之前曾是 active 状态的 NameNode 高。

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

- 如果二个 NameNode 都认为自己是 active 状态，则允许写入编辑日志的唯一 NameNode 是纪元号最新的那个 NameNode。纪元号较旧的 NameNode 将被禁止更改 HDFS 命名空间。

3. 设置高可用性的要求

在通过 QJM 方式配置 HDFS 高可用性前，你需要了解或准备以下事项：

- Standby NameNode 和 Primary NameNode 必须具有相同的硬件配置，包括 CPU、内存和磁盘大小。
- Cloudera 推荐你把 JournalNode 部署在元数据节点上（Primary NameNode、Standby NameNode、JobTracker 等）
- 为避免某个机架成为单点故障，高可用性配置中的任一节点（Primary NameNode、Standby NameNode、JobTracker 和 Backup JobTracker）不能和高可用性配置中的其他节点位于同一机架上。
- 你需要决定是否启用自动故障恢复。如果自动故障恢复被启用，你必须已在集群中安装了 ZooKeeper 组件，并指定一个 ZooKeeper quorum 来处理自动故障恢复。ZooKeeper 组必须包含奇数个节点，且必须包含至少三个节点。
- 如果高可用性被启用，则集群内不必有 secondary namenode 或其他检查点的服务。原因在于 Standby NameNode 可提供此类检查点的服务。
- 如果你有非高可用性集群且在 HDFS 中存有数据，则你需要决定在启用高可用性时是否保留这些 HDFS 数据。如果你需要进行这一操作，你需要知道如下限制：
    - 来自非高可用性集群的 DataNode 和 NameNode 不可从集群中删除。
    - 来自非高可用性集群的 DataNode 和 NameNode 在启用高可用性后，不可从之前所在机架中删除。
    - 来自非高可用性集群的 DataNode 的之前所在目录不可被删除。
- 如果你选择使用 Quorum Journal Manager（QJM），则你必须决定 QJM 集群中有多少节点。节点数量决定了集群是如何容错的。QJM 集群中允许失败且失败后可继续提供服务的节点数量根据以下算法计算： (N - 1) / 2

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

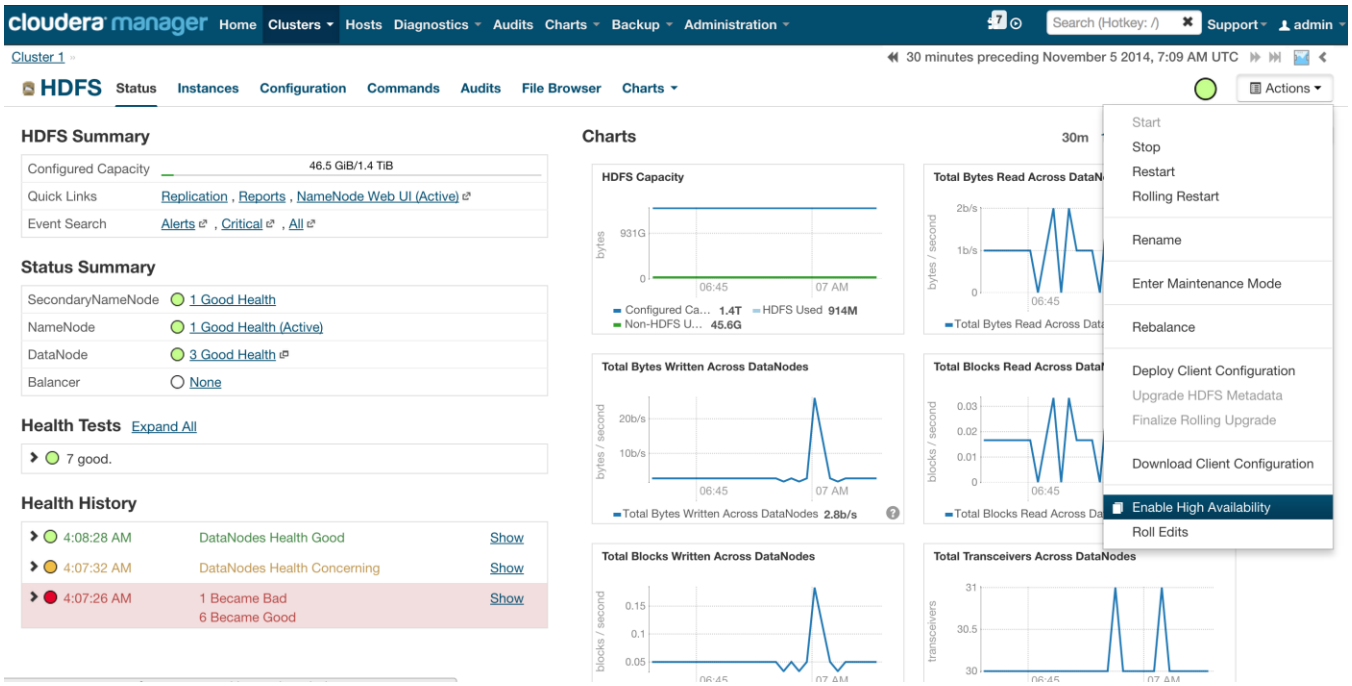
- 例如，如果集群中有 5 个 JournalNode，则允许失败的节点数量是 2，且失败后可继续提供服务。例如，如果集群中有 3 个 JournalNode，则允许失败且的节点数量是 1，且失败后可继续提供服务。

4. 设置高可用性

4.1 使用 QuorumJournalManager 进行共享存储

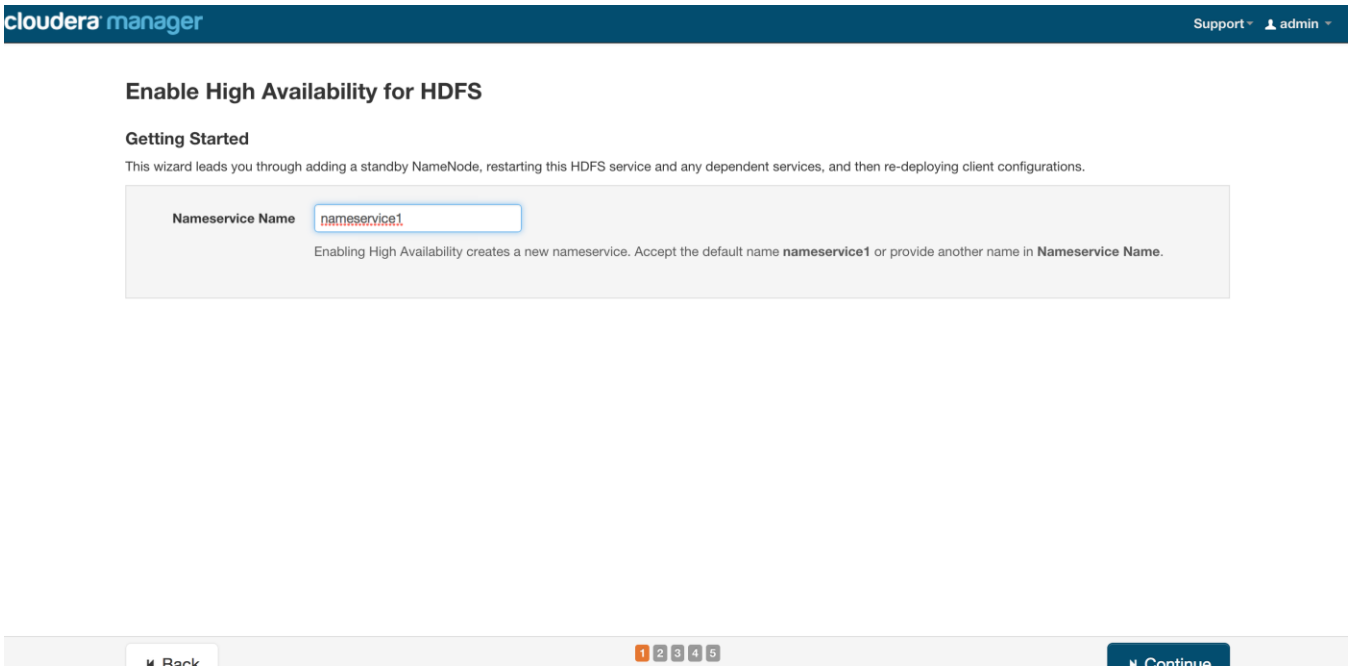
在 Cloudera Manager 5 中，HA 是通过 QJM 方式实现的。以下步骤演示了如何启用高可用性 HDFS，并启用自动故障恢复。

1. 在 Cloudera Manager 中，进入 HDFS Service
2. 点击 Actions > Enable High Availability，如下图所示



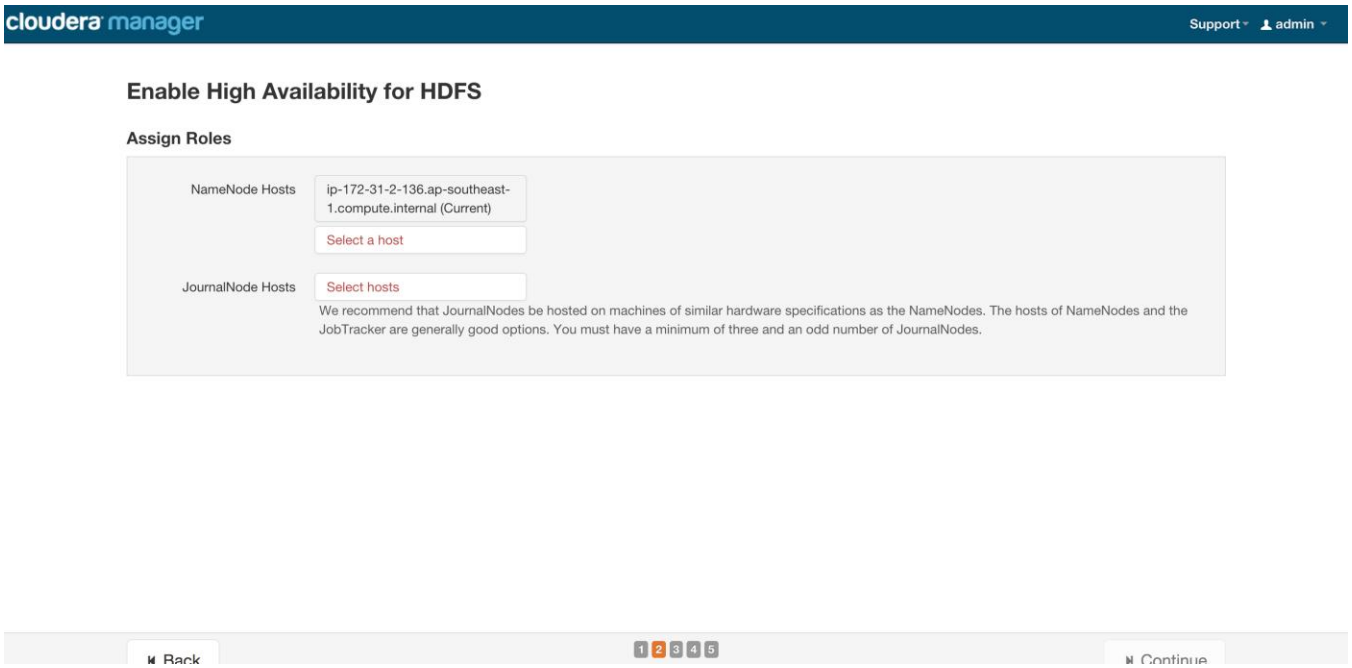
3. 指定一个 Nameservice 名称，默认为 nameservice1，点击 Continue 按钮继续

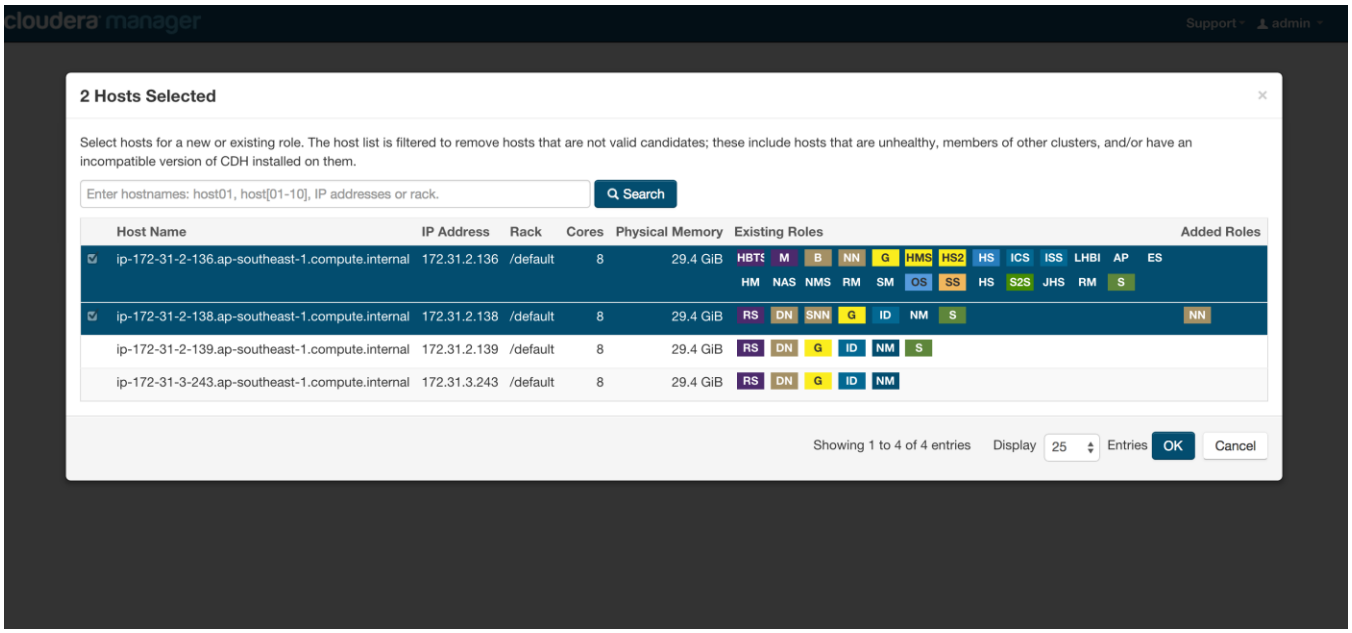
©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



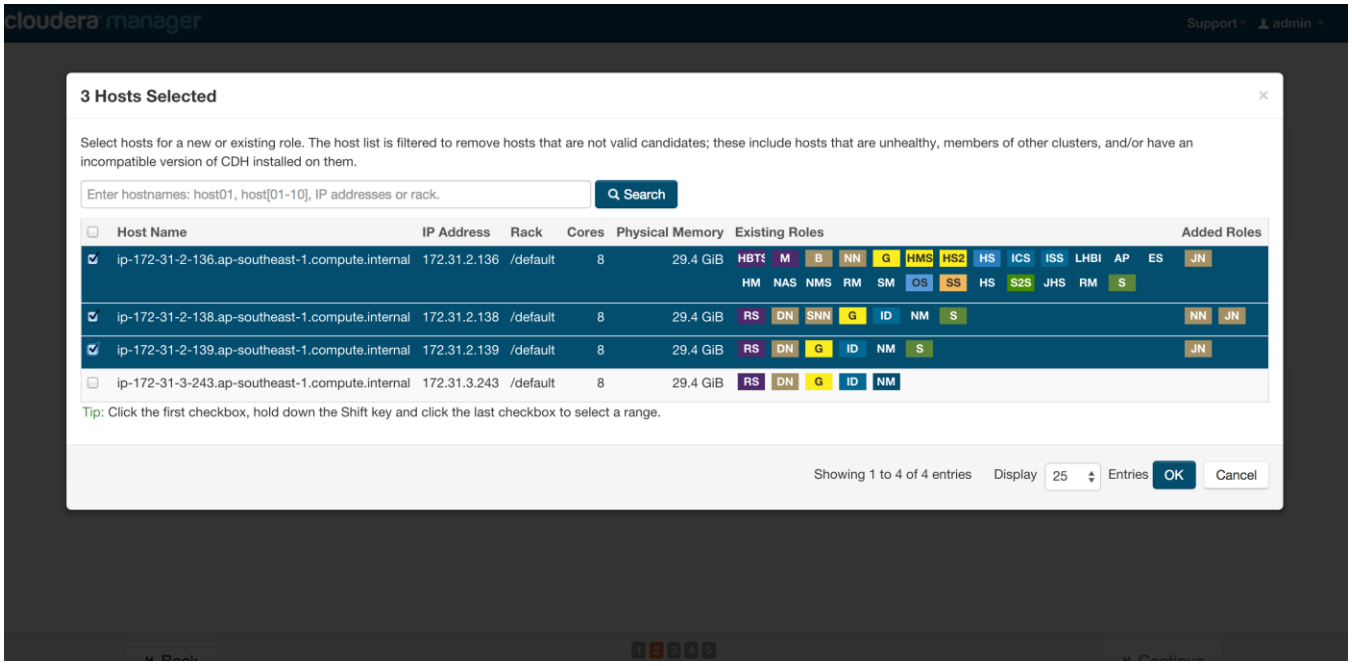
4. 在 NameNode Hosts 属性中, 点击 Select a host. 弹出选择主机窗口，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



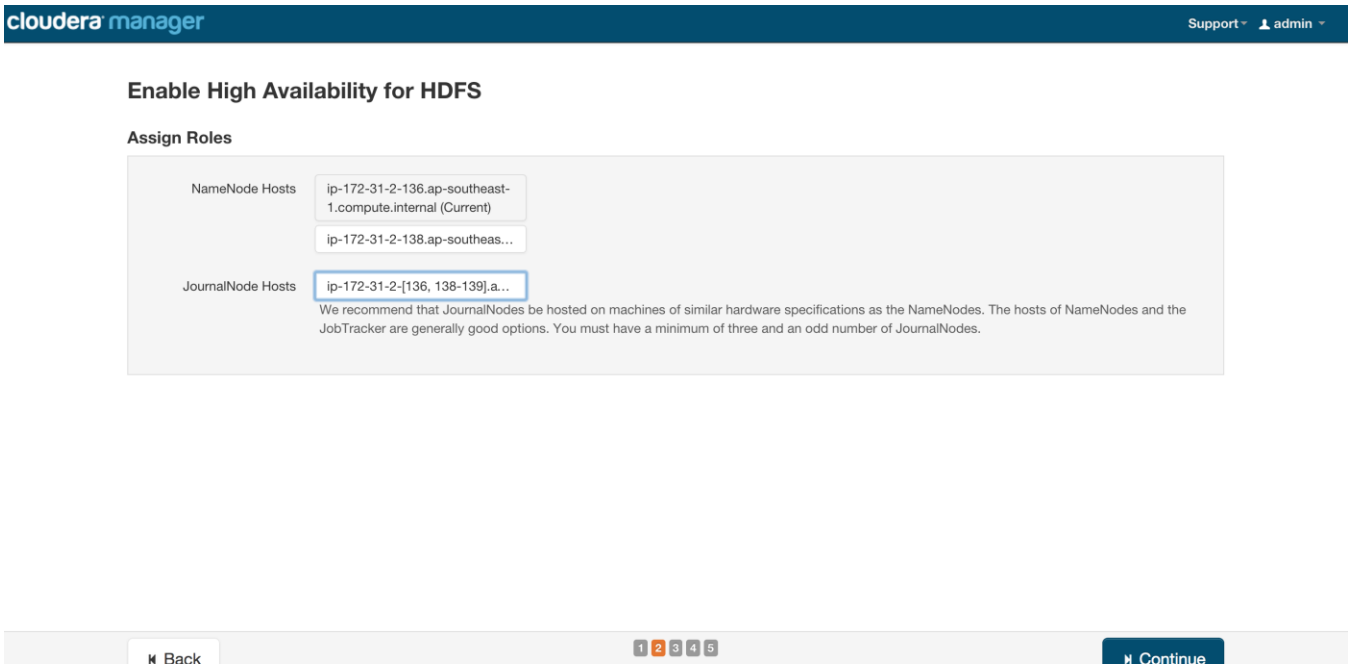


©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



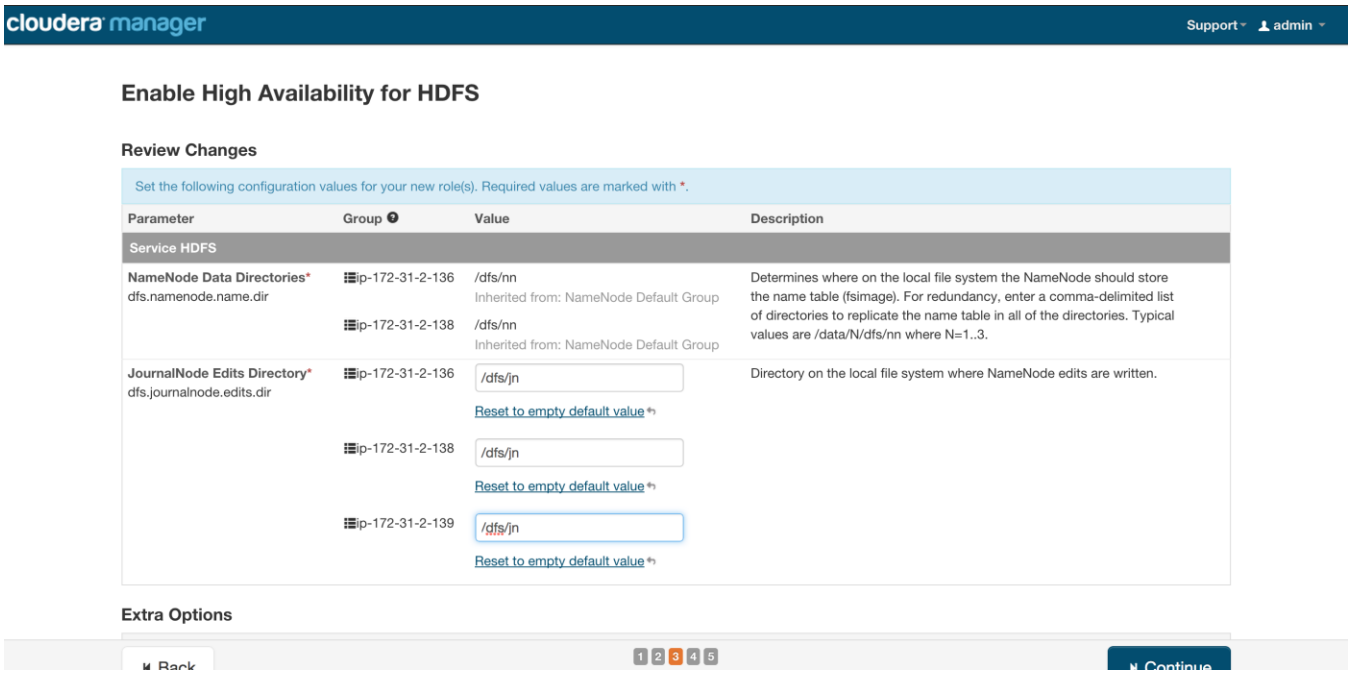
5. 指定 NameNode Hosts 和 JournalNode Hosts 后，点击 Continue 按钮继续

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



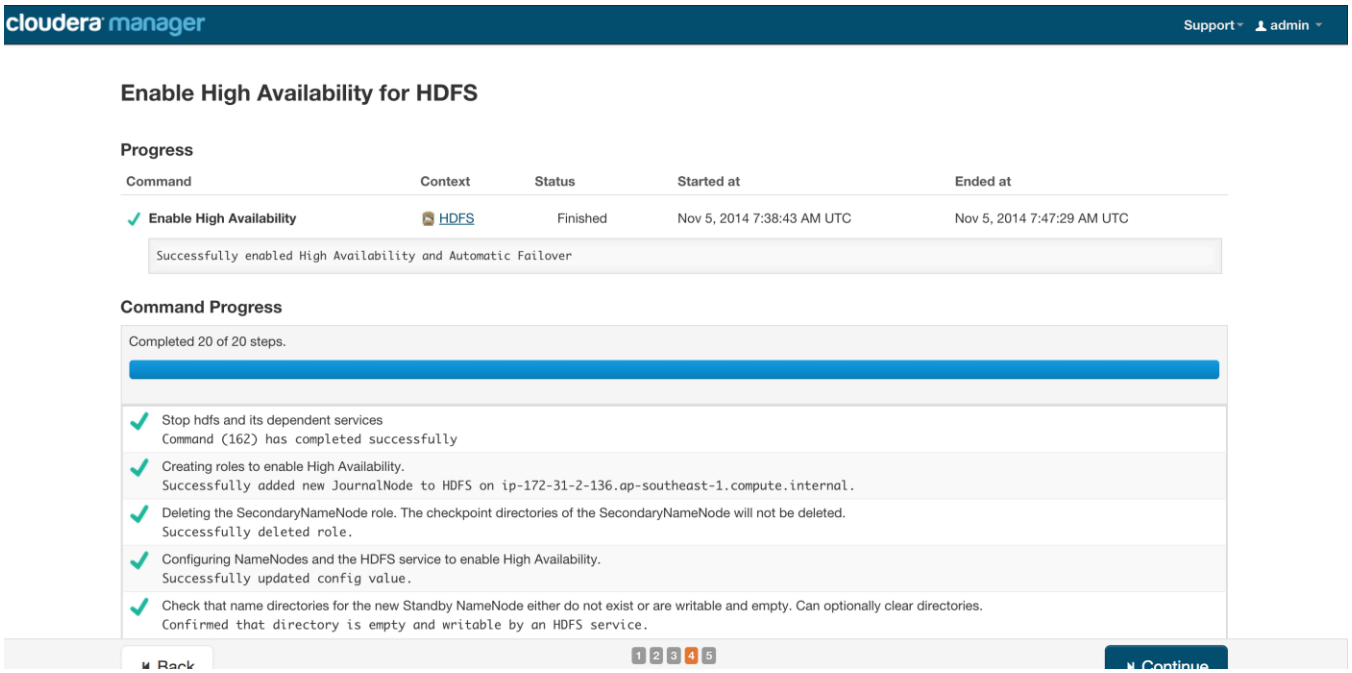
6. 填写 JournalNode Edits Directory，默认值为空，此处设为/dfs/jn，点击 Continue 按钮继续

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



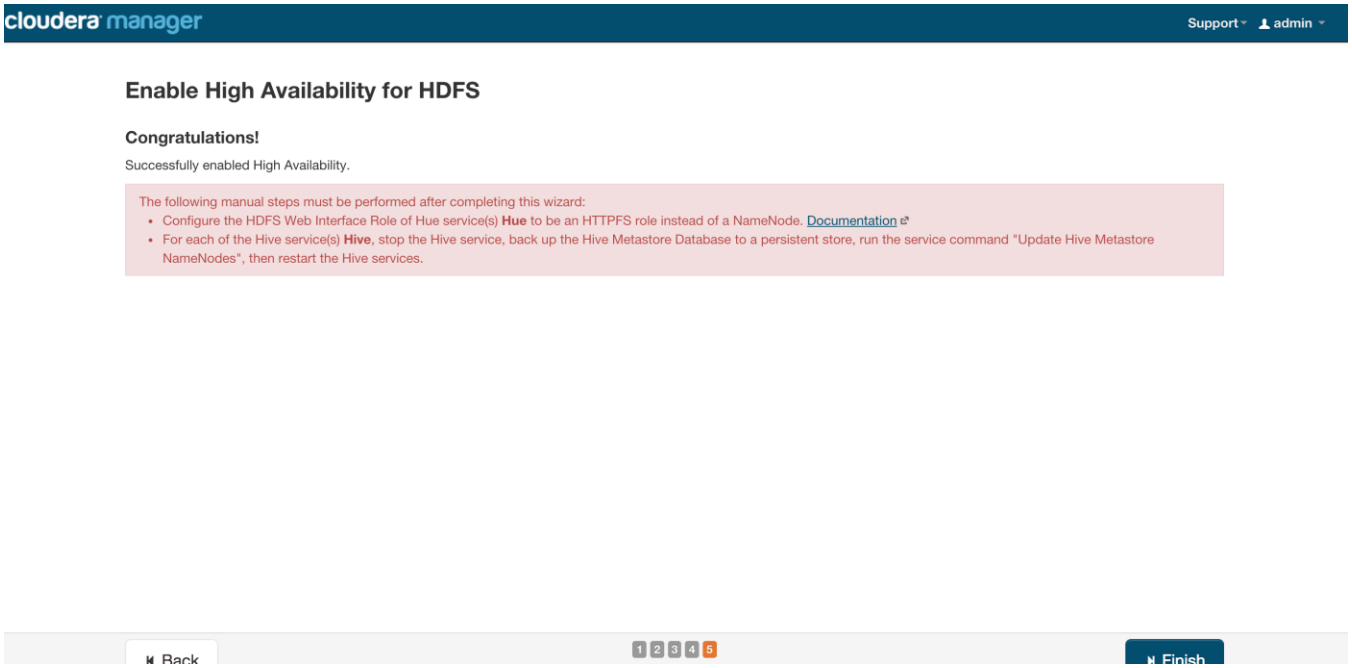
7. Cloudera Manager 执行一系列的命令进行配置，等待直到执行结束后，点击 Continue 按钮继续

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



8. 点击 Finish 按钮，HDFS 高可用性配置结束

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries.
All other trademarks are the property of their respective companies. Information is subject to change without notice.



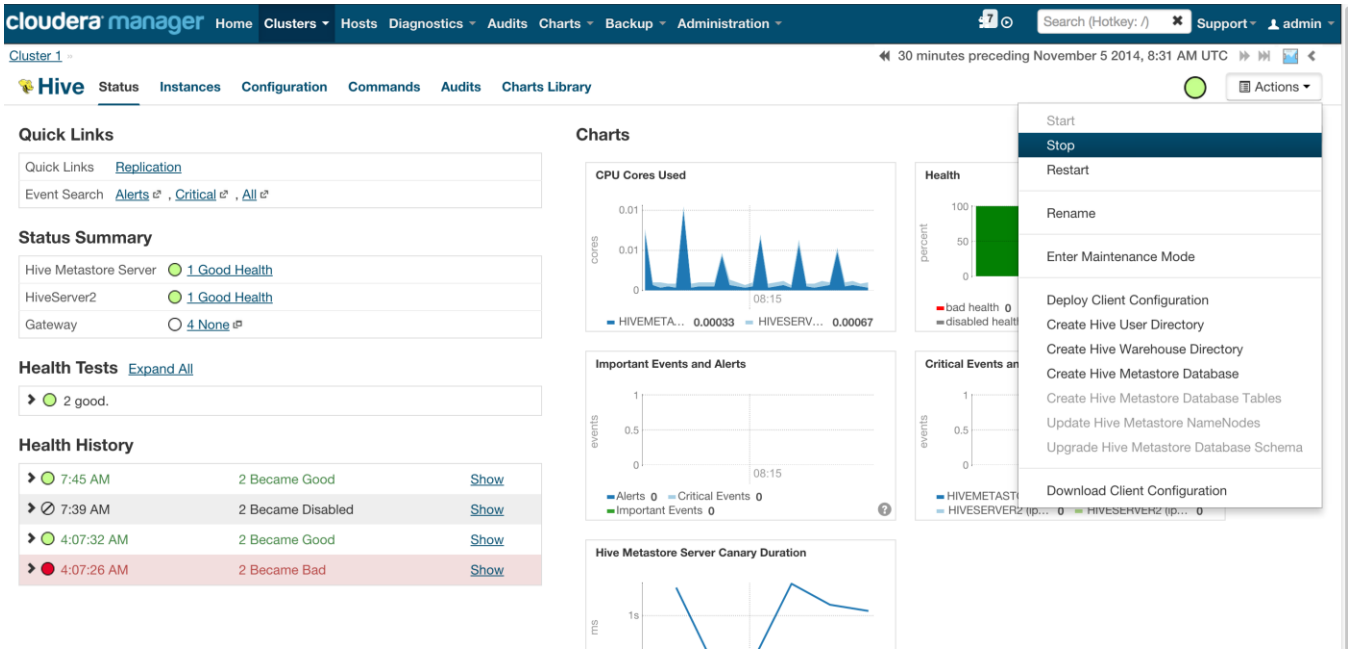
5. 配置 CDH 其他组件使用 HDFS 高可用性

配置 Hive Metastore、Hue、Impala 等 CDH 组件使用 HDFS 高可用性。

5.1 配置 Hive Metastore 使用 HDFS 高可用性

1. 在 Cloudera Manager 中，进入 Hive Service
2. 点击 Actions > Stop，如果 Hue 和 Impala 服务正在运行，需要先将其停止。如下图所示

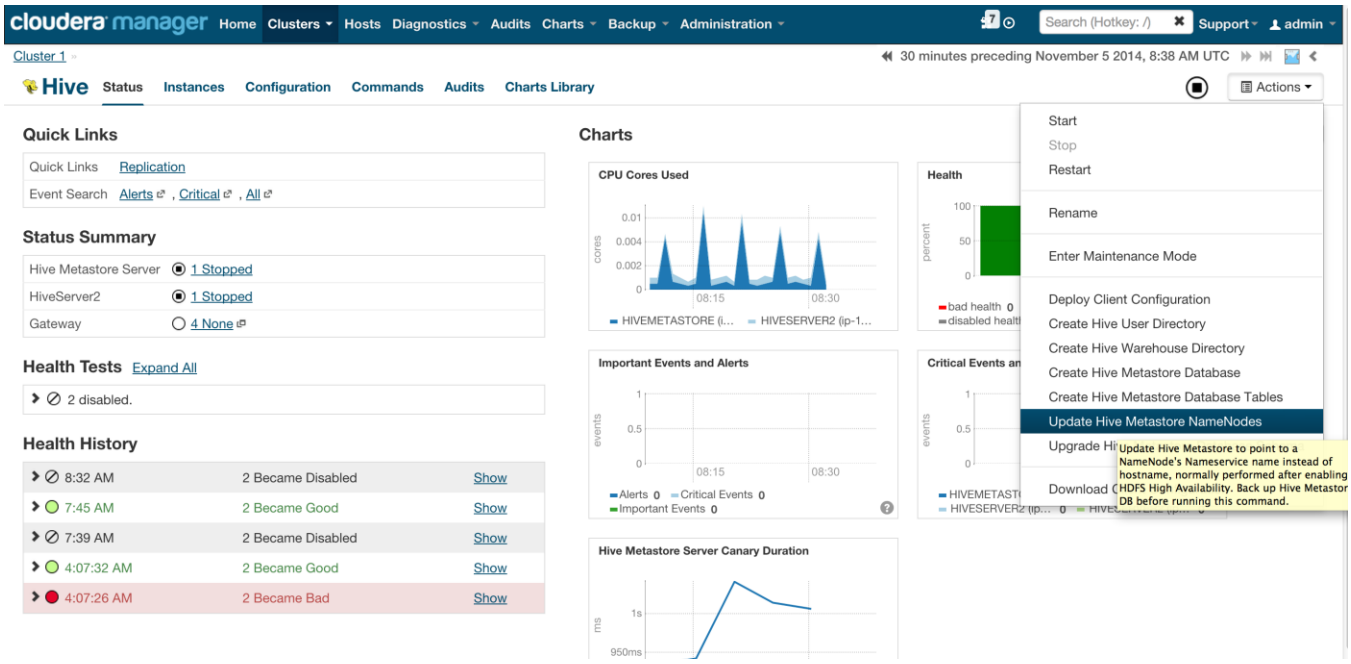
©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



3. 当 Hive 服务停止以后，请先备份 Hive Metastore 的数据，即将元数据从 MySQL（PostgreSQL 或 Oracle）库中导出到一个安全目录

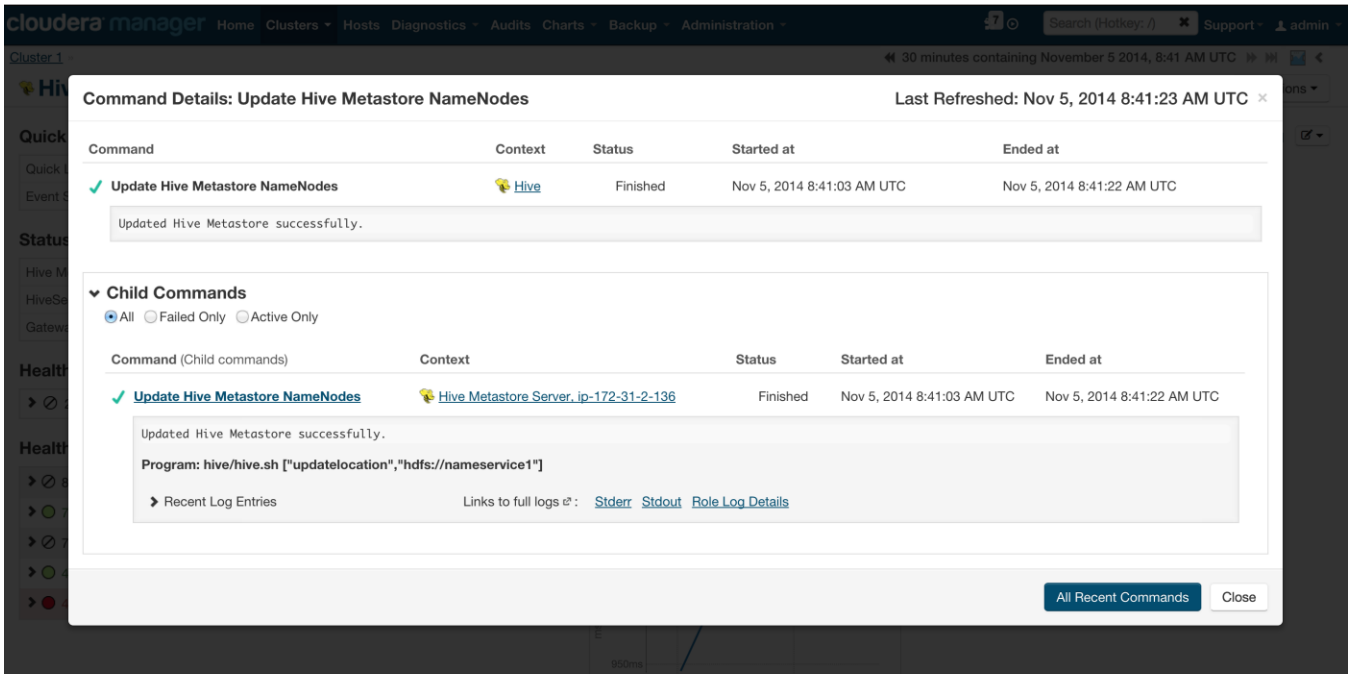
4. 选择 Actions >Update Hive Metastore NameNodes 并点击 Confirm 按钮确认。如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



5. 等待配置执行完毕后，关闭

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

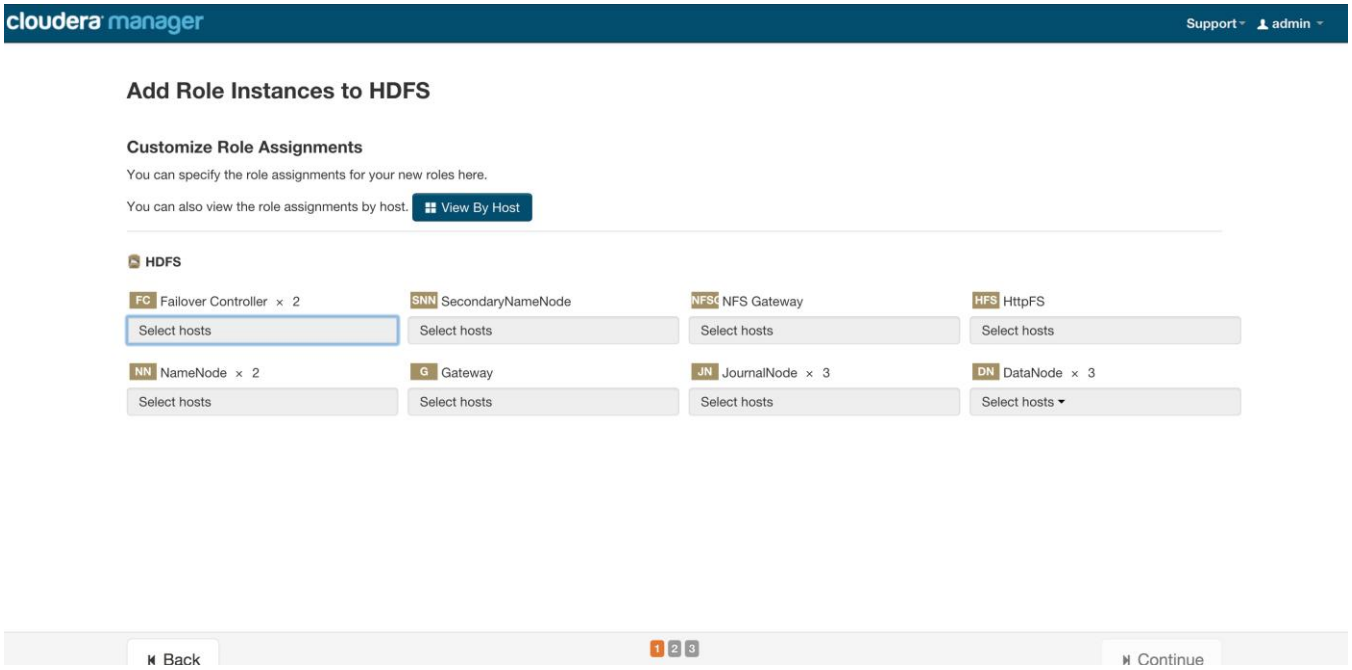


6. 点击 Actions > Start，如果 Hue 和 Impala 服务已停止，需要先将其启动

5.2 配置 Hue 使用 HDFS 高可用性

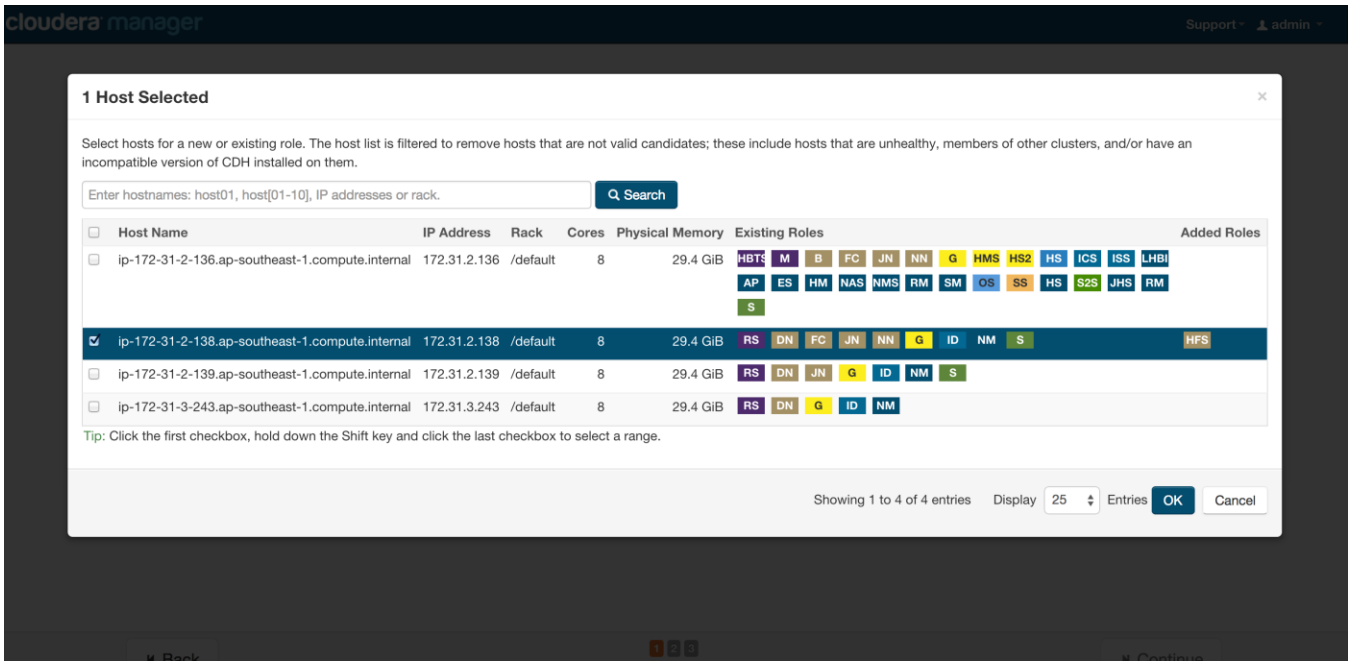
1. 在 Cloudera Manager 中，进入 HDFS Service
2. 进入 Instances 标签页面，点击 Add Role Instances 按钮，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries.
All other trademarks are the property of their respective companies. Information is subject to change without notice.



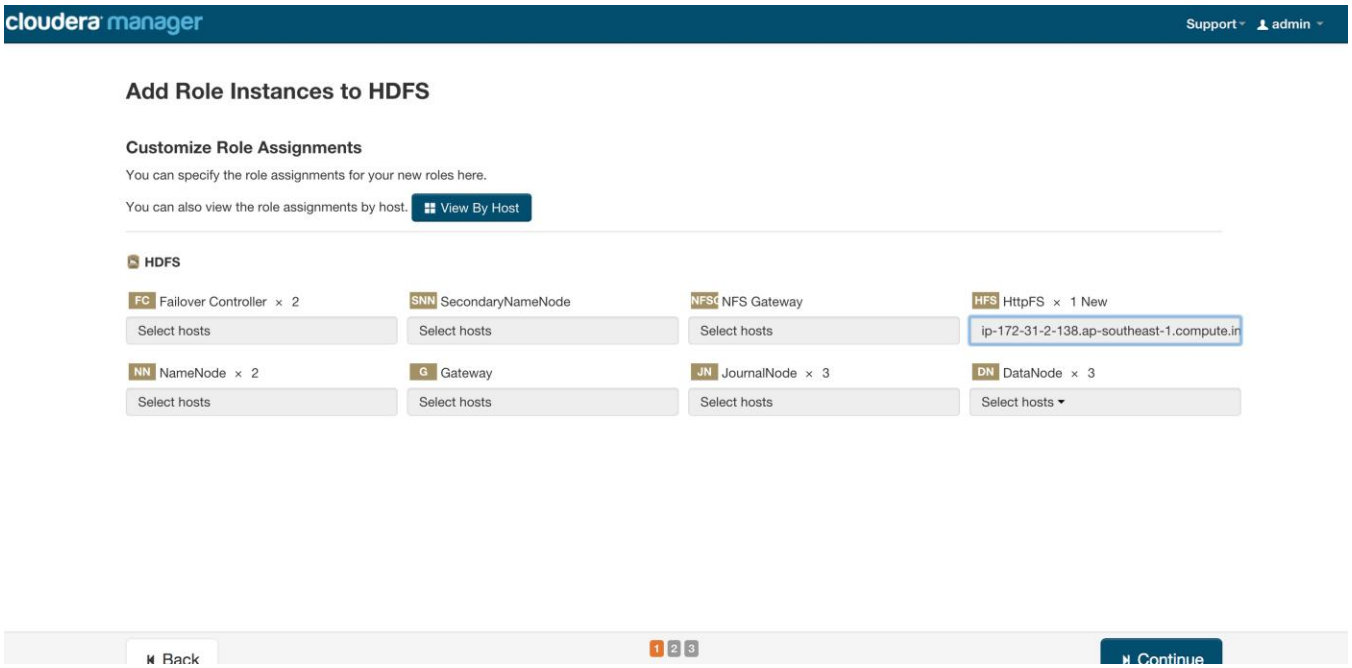
3. 点击 HttpFS 角色下面文本框选择主机，并点击 OK 按钮

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries.
All other trademarks are the property of their respective companies. Information is subject to change without notice.



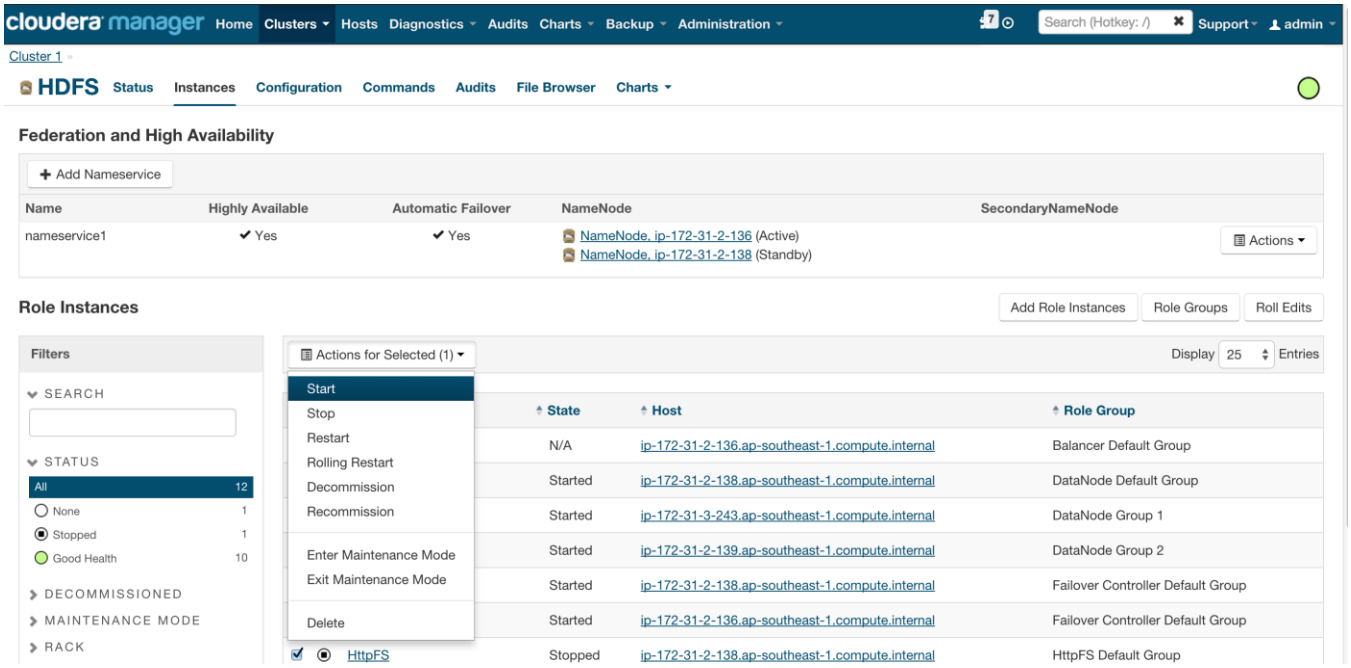
4. 点击 Continue 按钮，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.



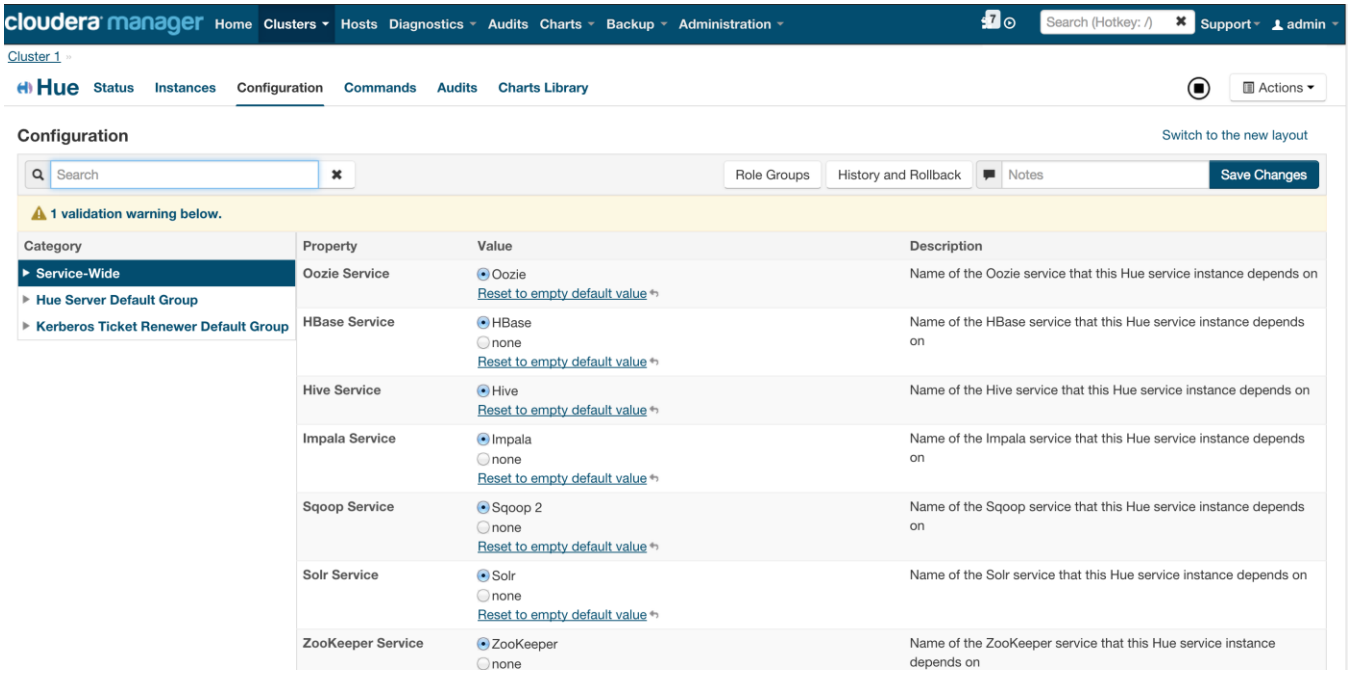
5. 返回 Instances 页面，选择 HttpFS 角色，并点击 Start 启动服务，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries.
All other trademarks are the property of their respective companies. Information is subject to change without notice.



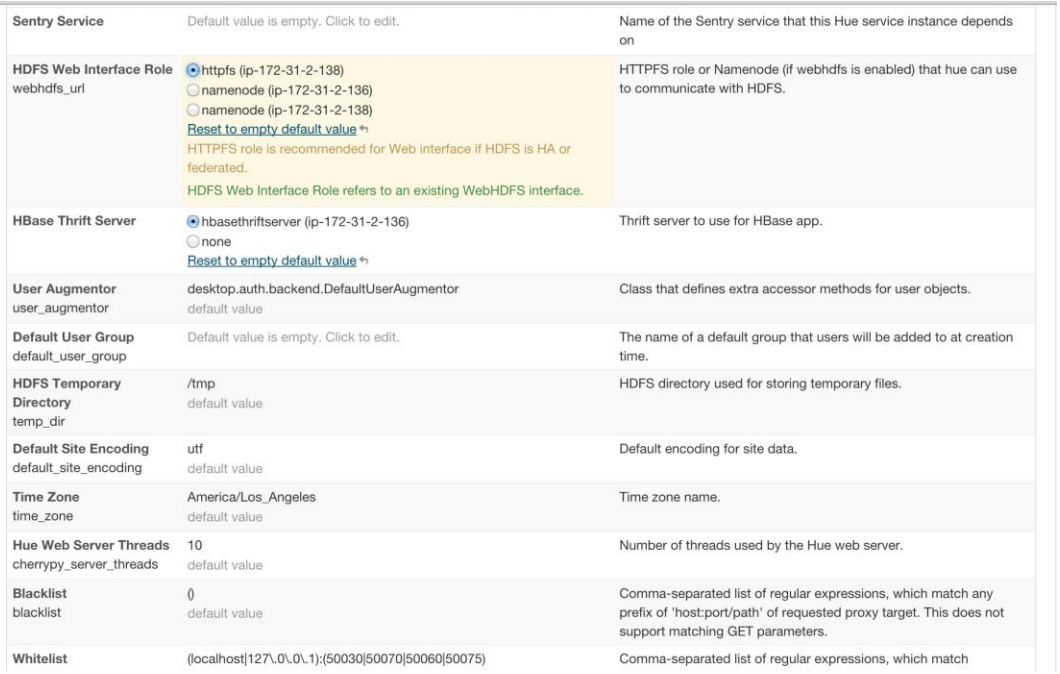
6. HttpFS 服务启动后，点击进入 Hue Service >Configuration 页面，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries.
All other trademarks are the property of their respective companies. Information is subject to change without notice.



7. 找到 Service-Wide > HDFS Web Interface Role 属性，选中 httpfs 单选框，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries.
All other trademarks are the property of their respective companies. Information is subject to change without notice.



Sentry Service	Default value is empty. Click to edit.	Name of the Sentry service that this Hue service instance depends on
HDFS Web Interface Role webhdfs_url	◉ httpfs (ip-172-31-2-138) ○ namenode (ip-172-31-2-136) ○ namenode (ip-172-31-2-138) Reset to empty default value HTTPFS role is recommended for Web interface if HDFS is HA or federated. HDFS Web Interface Role refers to an existing WebHDFS interface.	HTTPFS role or Namenode (if webhdfs is enabled) that hue can use to communicate with HDFS.
HBase Thrift Server	◉ hbasethriftserver (ip-172-31-2-136) ○ none Reset to empty default value	Thrift server to use for HBase app.
User Augmentor user_augmentor	desktop.auth.backend.DefaultUserAugmentor default value	Class that defines extra accessor methods for user objects.
Default User Group default_user_group	Default value is empty. Click to edit.	The name of a default group that users will be added to at creation time.
HDFS Temporary Directory temp_dir	/tmp default value	HDFS directory used for storing temporary files.
Default Site Encoding default_site_encoding	utf default value	Default encoding for site data.
Time Zone time_zone	America/Los_Angeles default value	Time zone name.
Hue Web Server Threads cherrypy_server_threads	10 default value	Number of threads used by the Hue web server.
Blacklist blacklist	() default value	Comma-separated list of regular expressions, which match any prefix of 'host:port/path' of requested proxy target. This does not support matching GET parameters.
Whitelist	(localhost|127\.0\.0\.1):(50030|50070|50060|50075)	Comma-separated list of regular expressions, which match

8. 点击 Save Changes 按钮保存修改并重启 Hue 服务

5.3 配置 Impala 使用 HDFS 高可用性

1. 确保之前的 5.1 章节 Hive Metastore 使用 HDFS 高可用性配置成功
2. 进入 impala shell，执行 INVALIDATE METADATA 命令，如下图所示

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.

```
[SSH] Server Version OpenSSH_5.3
[SSH] Logged in ()

Last login: Wed Nov  5 08:22:23 2014 from 74.217.76.11
[root@ip-172-31-3-243 ~]# impala
impalad       impala-shell
[root@ip-172-31-3-243 ~]# impala-shell
Starting Impala Shell without Kerberos authentication
Connected to ip-172-31-3-243.ap-southeast-1.compute.internal:21000
Server version: impalad version 2.0.0-cdh5 RELEASE (build ecf30af0b4d6e56ea80297df2189367ada6b7da7)
Welcome to the Impala shell. Press TAB twice to see a list of available commands.

Copyright (c) 2012 Cloudera, Inc. All rights reserved.

(Shell build version: Impala Shell v2.0.0-cdh5 (ecf30af) built on Sat Oct 11 13:56:06 PDT 2014)
[ip-172-31-3-243.ap-southeast-1.compute.internal:21000] > INVALIDATE METADATA;
Query: invalidate METADATA

Fetched 0 row(s) in 1.32s
[ip-172-31-3-243.ap-southeast-1.compute.internal:21000] >
```

6. 参考

[**1**] CDH HDFS High Availability

©2012 Cloudera, Inc. All rights reserved. Cloudera and the Cloudera logo are trademarks or registered trademarks of Cloudera Inc. in the USA and other countries. All other trademarks are the property of their respective companies. Information is subject to change without notice.