Hadoop分布式文件系统使用指南

目录

1．目的

本文档的目标是为Hadoop分布式文件系统（HDFS）的用户提供一个学习的起点，这里的HDFS既可以作为Hadoop集群的一部分，也可以作为一个独立的分布式文件系统。虽然HDFS在很多环境下被设计成是可正确工作的，但是了解HDFS的工作原理对在特定集群上改进HDFS的运行性能和错误诊断都有极大的帮助。

2．概述

HDFS是Hadoop应用用到的一个最主要的分布式存储系统。一个HDFS集群主要由一个NameNode和很多个Datanode组成：Namenode管理文件系统的元数据，而Datanode存储了实际的数据。HDFS的体系结构在这里有详细的描述。本文档主要关注用户以及管理员怎样和HDFS进行交互。HDFS架构设计中的图解描述了Namenode、Datanode和客户端之间的基本的交互操作。基本上，客户端联系Namenode以获取文件的元数据或修饰属性，而真正的文件I/O操作是直接和Datanode进行交互的。

下面列出了一些多数用户都比较感兴趣的重要特性。

- Hadoop（包括HDFS）非常适合在商用硬件（commodity hardware）上做分布式存储和计算，因为它不仅具有容错性和可扩展性，而且非常易于扩展。Map-Reduce框架以其在大型分布式系统应用上的简单性和可用性而著称，这个框架已经被集成进Hadoop中。
- HDFS的可配置性极高，同时，它的默认配置能够满足很多的安装环境。多数情况下，这些参数只在非常大规模的集群环境下才需要调整。
- 用Java语言开发，支持所有的主流平台。
- 支持类She11命令，可直接和HDFS进行交互。
- NameNode和DataNode有内置的Web服务器，方便用户检查集群的当前状态。
- 新特性和改进会定期加入HDFS的实现中。下面列出的是HDFS中常用特性的一部分：
  - 文件权限和授权。
  - 机架感知（Rack awareness）：在调度任务和分配存储空间时考虑节点的物理位置。
  - 安全模式：一种维护需要的管理模式。
  - fsck：一个诊断文件系统健康状况的工具，能够发现丢失的文件或数据块。
  - Rebalancer：当datanode之间数据不均衡时，平衡集群上的数据负载。
  - 升级和回滚：在软件更新后有异常发生的情形下，能够回滚到HDFS升级之前的状态。
  - Secondary Namenode：对文件系统名字空间执行周期性的检查点，将Namenode

上HDFS改动日志文件的大小控制在某个特定的限度下。

3. 先决条件

下面的文档描述了如何安装和搭建Hadoop集群:

- Hadoop快速入门 针对初次使用者。
- Hadoop集群搭建 针对大规模分布式集群的搭建。

文档余下部分假设用户已经安装并运行了至少包含一个Datanode节点的HDFS。就本文目的来说, Namenode和Datanode可以运行在同一个物理主机上。

4. Web接口

NameNode和DataNode各自启动了一个内置的Web服务器, 显示了集群当前的基本状态和信息。在默认配置下NameNode的首页地址是http://namenode-name:50070/。这个页面列出了集群里的所有DataNode和集群的基本状态。这个Web接口也可以用来浏览整个文件系统（使用NameNode首页上的"Browse the file system"链接）。

5. Shell命令

Hadoop包括一系列的类shell的命令, 可直接和HDFS以及其他Hadoop支持的文件系统进行交互。bin/hadoop fs -help 命令列出所有Hadoop Shell支持的命令。而 bin/hadoop fs -help command-name 命令能显示关于某个命令的详细信息。这些命令支持大多数普通文件系统的操作, 比如复制文件、改变文件权限等。它还支持一些HDFS特有的操作, 比如改变文件副本数目。

5.1. DFSAdmin命令

'bin/hadoop dfsadmin' 命令支持一些和HDFS管理相关的操作。bin/hadoop dfsadmin -help 命令能列出所有当前支持的命令。比如:

- -report: 报告HDFS的基本统计信息。有些信息也可以在NameNode Web服务首页看到。
- -safemode: 虽然通常并不需要, 但是管理员的确可以手动让NameNode进入或离开安全模式。
- -finalizeUpgrade: 删除上一次升级时制作的集群备份。

6. Secondary NameNode

NameNode将对文件系统的改动追加保存到本地文件系统上的一个日志文件（edits）。当一个NameNode启动时，它首先从一个映像文件（fsimage）中读取HDFS的状态，接着应用日志文件中的edits操作。然后它将新的HDFS状态写入（fsimage）中，并使用一个空的edits文件开始正常操作。因为NameNode只有在启动阶段才合并fsimage和edits，所以久而久之日志文件可能会变得非常庞大，特别是对大型的集群。日志文件太大的另一个副作用是下一次NameNode启动会花很长时间。

Secondary NameNode定期合并fsimage和edits日志，将edits日志文件大小控制在一个限度下。因为内存需求和NameNode在一个数量级上，所以通常secondary NameNode和NameNode运行在不同的机器上。Secondary NameNode通过bin/start-dfs.sh在conf/masters中指定的节点上启动。

Secondary NameNode的检查点进程启动，是由两个配置参数控制的：

- fs.checkpoint.period，指定连续两次检查点的最大时间间隔， 默认值是1小时。
- fs.checkpoint.size定义了edits日志文件的最大值，一旦超过这个值会导致强制执行检查点（即使没到检查点的最大时间间隔）。默认值是64MB。

Secondary NameNode保存最新检查点的目录与NameNode的目录结构相同。 所以NameNode可以在需要的时候读取Secondary NameNode上的检查点镜像。

如果NameNode上除了最新的检查点以外，所有的其他的历史镜像和edits文件都丢失了， NameNode可以引入这个最新的检查点。以下操作可以实现这个功能：

- 在配置参数dfs.name.dir指定的位置建立一个空文件夹；
- 把检查点目录的位置赋值给配置参数fs.checkpoint.dir；
- 启动NameNode，并加上-importCheckpoint。

NameNode会从fs.checkpoint.dir目录读取检查点， 并把它保存在dfs.name.dir目录下。 如果dfs.name.dir目录下有合法的镜像文件，NameNode会启动失败。 NameNode会检查fs.checkpoint.dir目录下镜像文件的一致性，但是不会去改动它。

命令的使用方法请参考secondarynamenode 命令.

7. Rebalancer

HDFS的数据也许并不是非常均匀的分布在各个DataNode中。一个常见的原因是在现有

的集群上经常会增添新的DataNode节点。当新增一个数据块（一个文件的数据被保存在一系列的块中）时，NameNode在选择DataNode接收这个数据块之前，会考虑到很多因素。其中的一些考虑的是：

- 将数据块的一个副本放在正在写这个数据块的节点上。
- 尽量将数据块的不同副本分布在不同的机架上，这样集群可在完全失去某一机架的情况下还能存活。
- 一个副本通常被放置在和写文件的节点同一机架的某个节点上，这样可以减少跨越机架的网络I/O。
- 尽量均匀地将HDFS数据分布在集群的DataNode中。

由于上述多种考虑需要取舍，数据可能并不会均匀分布在DataNode中。HDFS为管理员提供了一个工具，用于分析数据块分布和重新平衡DataNode上的数据分布。HADOOP-1652的附件中的一个PDF是一个简要的rebalancer管理员指南。

使用方法请参考balancer 命令.

8．机架感知（Rack awareness）

通常，大型Hadoop集群是以机架的形式来组织的，同一个机架上不同节点间的网络状况比不同机架之间的更为理想。另外，NameNode设法将数据块副本保存在不同的机架上以提高容错性。Hadoop允许集群的管理员通过配置dfs.network.script参数来确定节点所处的机架。当这个脚本配置完毕，每个节点都会运行这个脚本来获取它的机架ID。默认的安装假定所有的节点属于同一个机架。这个特性及其配置参数在HADOOP-692所附的PDF上有更详细的描述。

9．安全模式

NameNode启动时会从fsimage和edits日志文件中装载文件系统的状态信息，接着它等待各个DataNode向它报告它们各自的数据块状态，这样，NameNode就不会过早地开始复制数据块，即使在副本充足的情况下。这个阶段，NameNode处于安全模式下。NameNode的安全模式本质上是HDFS集群的一种只读模式，此时集群不允许任何对文件系统或者数据块修改的操作。通常NameNode会在开始阶段自动地退出安全模式。如果需要，你也可以通过'bin/hadoop dfsadmin -safemode'命令显式地将HDFS置于安全模式。NameNode首页会显示当前是否处于安全模式。关于安全模式的更多介绍和配置信息请参考JavaDoc：setSafeMode()。

10. fsck

HDFS支持fsck命令来检查系统中的各种不一致状况。这个命令被设计来报告各种文件存在的问题，比如文件缺少数据块或者副本数目不够。不同于在本地文件系统上传统的fsck工具，这个命令并不会修正它检测到的错误。一般来说，NameNode会自动修正大多数可恢复的错误。HDFS的fsck不是一个Hadoop shell命令。它通过'bin/hadoop fsck'执行。 命令的使用方法请参考fsck命令 fsck可用来检查整个文件系统，也可以只检查部分文件。

11. 升级和回滚

当在一个已有集群上升级Hadoop时，像其他的软件升级一样，可能会有新的bug或一些会影响到现有应用的非兼容性变更出现。在任何有实际意义的HDSF系统上，丢失数据是不被允许的，更不用说重新搭建启动HDFS了。HDFS允许管理员退回到之前的Hadoop版本，并将集群的状态回滚到升级之前。更多关于HDFS升级的细节在升级wiki上可以找到。HDFS在一个时间可以有一个这样的备份。在升级之前，管理员需要用bin/hadoop dfsadmin -finalizeUpgrade（升级终结操作）命令删除存在的备份文件。下面简单介绍一下一般的升级过程:

- 升级 Hadoop 软件之前，请检查是否已经存在一个备份，如果存在，可执行升级终结操作删除这个备份。通过dfsadmin -upgradeProgress status命令能够知道是否需要对一个集群执行升级终结操作。
- 停止集群并部署新版本的Hadoop。
- 使用-upgrade选项运行新的版本（bin/start-dfs.sh -upgrade）。
- 在大多数情况下，集群都能够正常运行。一旦我们认为新的HDFS运行正常（也许经过几天的操作之后），就可以对之执行升级终结操作。注意，在对一个集群执行升级终结操作之前，删除那些升级前就已经存在的文件并不会真正地释放DataNodes上的磁盘空间。
- 如果需要退回到老版本，
  - 停止集群并且部署老版本的Hadoop。
  - 用回滚选项启动集群（bin/start-dfs.h -rollback）。

12. 文件权限和安全性

这里的文件权限和其他常见平台如Linux的文件权限类似。目前，安全性仅限于简单的文件权限。启动NameNode的用户被视为HDFS的超级用户。HDFS以后的版本将会支持网

络验证协议（比如Kerberos）来对用户身份进行验证和对数据进行加密传输。具体的细节请参考权限使用管理指南。

13．可扩展性

现在，Hadoop已经运行在上千个节点的集群上。Powered By Hadoop页面列出了一些已将Hadoop部署在他们的大型集群上的组织。HDFS集群只有一个NameNode节点。目前，NameNode上可用内存大小是一个主要的扩展限制。在超大型的集群中，增大HDFS存储文件的平均大小能够增大集群的规模，而不需要增加NameNode的内存。默认配置也许并不适合超大规模的集群。Hadoop FAQ页面列举了针对大型Hadoop集群的配置改进。

14．相关文档

这个用户手册给用户提供了一个学习和使用HDSF文件系统的起点。本文档会不断地进行改进，同时，用户也可以参考更多的Hadoop和HDFS文档。下面的列表是用户继续学习的起点：

- Hadoop官方主页：所有Hadoop相关的起始页。
- Hadoop Wiki：Hadoop Wiki文档首页。这个指南是Hadoop代码树中的一部分，与此不同，Hadoop Wiki是由Hadoop社区定期编辑的。
- Hadoop Wiki上的FAQ。
- Hadoop JavaDoc API。
- Hadoop用户邮件列表：core-user[at]hadoop.apache.org。
- 查看conf/hadoop-default.xml文件。这里包括了大多数配置参数的简要描述。
- 命令手册：命令使用说明。