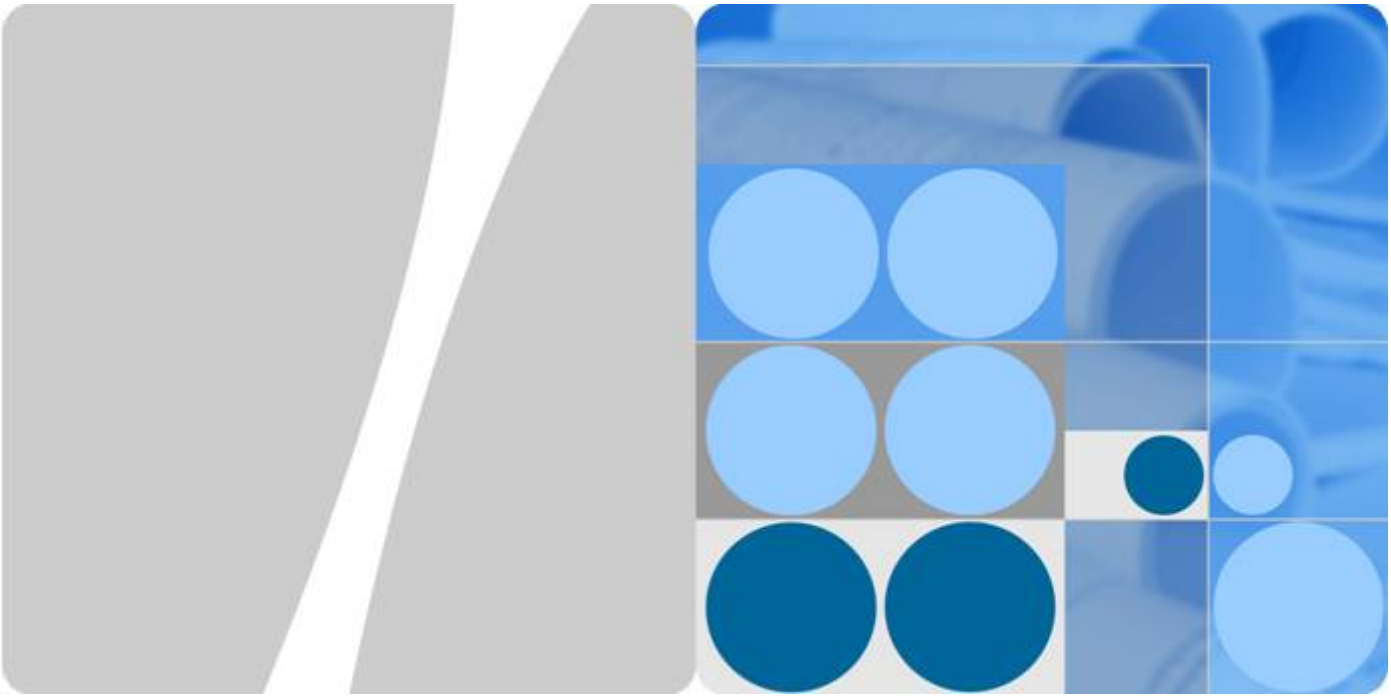


FusionInsight HD 性能调优指导手册 - HDFS

FusionInsight HD 性能调优指导手册 - HDFS

文档版本 01
发布日期 2016-12-16

华为技术有限公司

HUAWEI

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

目 录

FusionInsight HD 性能调优指导手册 - HDFS

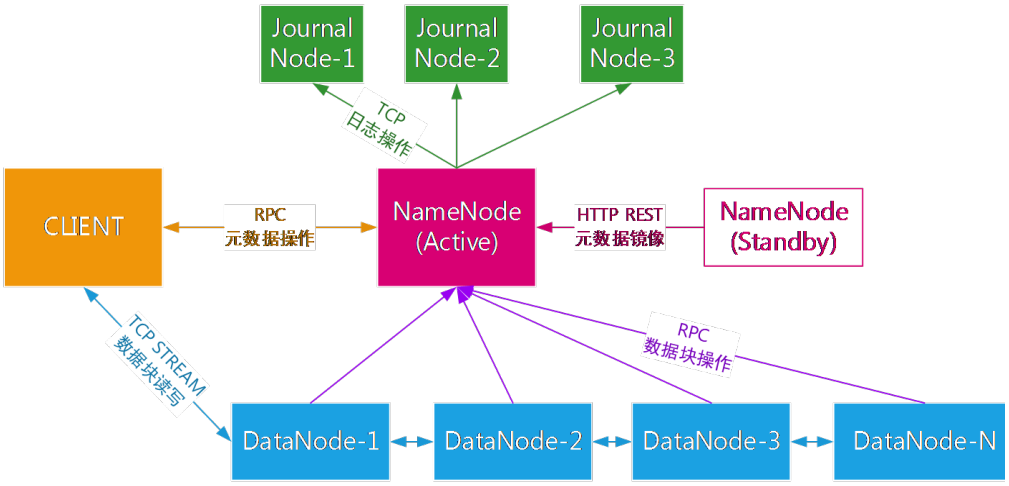
FusionInsight HD 性能调优指导手册 - HDFS

1 概述

FusionInsight HD 性能调优指导手册 - HDFS

1.1 模块架构模型

图 1-1 HDFS 架构



HDFS采用master/slave架构。一个HDFS集群是由一对Namenode、三个JournalNode，一定数目的Datanodes，多个客户端组成。

HDFS架构主要包括：

- Namenode：执行文件系统的名字空间操作，比如打开、关闭、重命名文件或目录。它也负责确定数据块到具体Datanode节点的映射。
- Datanode：负责处理文件系统客户端的读写请求。在Namenode的统一调度下进行数据块的创建、删除和复制。
- Client：支持业务访问HDFS，从NameNode ,DataNode获取数据返回给业务。多个实例，和业务一起运行。
- JournalNode：文件系统共享EditLog服务器，用于存储EditLog，由三个节点组成。

通信方式：

- Client与NameNode采用RPC方式通信，消息内容为元数据类操作：打开、关闭、重命名文件，目录的创建、删除等。
- Client与DataNode采用TCP流方式通信，消息内容为数据块。
- NameNode与DataNode采用RPC方式通信，消息内容为数据块状态上报，以及操作命令字。

说明

基于此架构方式，HDFS适用于高吞吐量，而不适合低时间延迟的访问（因为写文件可能存在初始化socket，初始化RPC，及其多次namenode通信等）。如果对于一个大文件，它写数据流的时间远远大于RPC通信，socket建立连接，及其磁盘寻址等时间，因此大文件比较适合；如果对于小文件则相反，其他方面所消耗时间比写数据流的时间消耗得多。

对于查询文件，创建目录，设置文件权限等操作则关注NameNode相关的性能处理。如果涉及读写文件吞吐量的相关的问题主要是基于Client和DataNode进行优化。对于小文件大量的读取也需要关注NameNode的处理能力。

FusionInsight HD 性能调优指导手册 - HDFS

1.2 性能衡量指标

1.2.1 衡量指标

HDFS属于存储系统，存储系统衡量指标有两个OPS和Throughput.

- **OPS**

每秒操作次数，即每秒元数据的操作次数，反应的是NameNode的处理性能，即Client->NameNode的处理能力，NameNode的写入能力，受到JournalNode的影响。

受影响的API列表：

API	说明
Open	文件打开
Create	创建文件
Close	关闭文件
Mkdirs	创建文件
Rmdir	删除文件。
Setacl	设置权限
Removeacl	删除权限
Setowner	设置属组
Setpermission	设置权限
Getstatus	获取文件属性
Liststatus	获取目录文件列表。

- **Throughput**

即吞吐量，处理数据块的读写速度，反应的是DataNode的处理能力，即

Client->DataNode的处理能力。

受影响的API列表：

API	说明
Inputstream.read()	读取文件
Outputstream.write	写文件
Outputstream.flush()	强制同步文件
Outputstream.close()	同步并关闭文件

- **Capacity**

即容量，反应的是HDFS存储的最大文件数目和存储空间容量。

1.2.2 指标观测方法

- Capacity

指标名	指标说明	指标意义
HDFS File	集群中文件数目	● NameNode 启动时间 ● NameNode 内存占用大小
Total Blocks	集群中数据块数目	● NameNode 启动时间 ● NameNode 内存占用大小
Blocks	DataNode 数据块个数	● DataNode启动时间 ● 内存占用大小

说明

NameNode的处理能力可以通过调整日志处理等一些特性实现。或者通过启用Federation特性新增NameNode。以上指标，可以观察NameNode 的容量情况，和消息处理能力

HDFS File + Total Blocks 之后建议小于3亿，大概对应1.5亿小文件左右

- OPS

指标名	指标说明	指标意义
Average RPC Queueing Time	RPC平均等待时间	NameNode处理压力
Average RPC Queueing Time	RPC平均处理时间。	Client单次请求处理时间。
File Creation Operations		
Rename Operations		
GetBlockLocations Operations		
Mkdirs Operations		
GetFileInfo Operations		
File Read Operations		

说明

Average RPC Queuing Time 和 Average RPC Processing Time 两个值 正常都在20ms以内。当前HDFS默认支持的OPS，全写大概3K，全读在11K左右，通常根据**Average RPC Queuing Time**和**Average RPC Processing Time**时长可以观测到NameNode对应的消息处理能力。如果需要关注业务具体的OPS也可以通过HDFS Service级别的以下监控项查看对应的业务请求，以做后续调优的根据

FusionInsight HD 性能调优指导手册 - HDFS

- Throughput

指标名	指标说明	指标意义
DataNode Read Write Bytes	DataNode读写字节数	DataNode读写字节数
Disk IO Speed of the DataNode	DataNode节点硬盘速度	DataNode节点硬盘速度
DataNode Operation Time	DataNode操作时间	DataNode操作时间

说明

DataNode的。

FusionInsight HD 性能调优指导手册 - HDFS

2 集群服务部署规划

FusionInsight HD 性能调优指导手册 - HDFS

2.1 规格能力

HDFS当前支持的最大规格能力如下:

Capacity

指标名	最大规格值	说明
HDFS File	150,000,000	一对主备NameNode最多支持的文件数量
Total Blocks	150,000,000	一对主备NameNode最多支持的Block数量，一个文件至少对应一个Block
Blocks	5,000,000	单个DataNode 最大支持的Block个数，由于集群是3副本，DataNode的节点个数需要 3* Total Blocks/Block

OPS

指标名	最大规格值	说明
OPS（write）	3K/s	NameNode只处理写操作的能力
OPS（read）	11K/s	NameNode只处理读操作的处理能力

Throughput

指标名	最大规格值	说明
Write speed	150M/s	单DataNode的写入能力，以10个SATA盘为例，在网络不是瓶颈的情况下, 450M/s左右。考虑默认3副本的情况
ReadSpeed	450M/s	DataNode节点硬盘速度

2.2 环境要求

服务器配置要求

以下描述NameNode和JournalNode属于控制节点，DataNode属于数据节点。

FusionInsight HD 性能调优指导手册 - HDFS

NameNode由主备两个节点组成，JournalNode由三个节点组成，DataNode根据业务需求配置，常见推荐配置如下表

2.2.1 常见配置

100,000,000 文件数量配置

类型	节点数	CPU	内存	硬盘	网络	说明
NameNode	2	8 Cores	≥128 GB	RAID1 SAS ≥600GB	2*10Gb光口	内存和文件对象呈线性比例关系，硬盘主要是持久化元数据和EditLog
JournalNode	3	8 Cores	2 GB	RAID1 SAS ≥600GB	2*10Gb光口	硬盘主要是持久化元数据和EditLog
DataNode	≥300	8 Cores	16 GB	SAS/SATA *N	2*10Gb光口	硬盘用于存储业务数据的blocks，通常是10块以上的4T SATA。 节点内存主要用于共部署的NodeManager任务使用

50,000,000 文件数量配置

类型	节点数	CPU	内存	硬盘	网络	说明
NameNode	2	8 Cores	≥128 GB	RAID1 SAS ≥600GB	2*10Gb光口	内存和文件对象呈线性比例关系，硬盘主要是持久化元数据和EditLog
JournalNode	3	8 Cores	2 GB	RAID1 SAS ≥600GB	2*10Gb光口	硬盘主要是持久化元数据和EditLog

FusionInsight HD 性能调优指导手册 - HDFS

类型	节点数	CPU	内存	硬盘	网络	说明
DataNode	≥300	8 Cores	16 GB	SAS/SATA *N	2*10Gb光口	硬盘用于存储业务数据的blocks，通常是10块以上的4T SATA。 节点内存主要用于共部署的NodeManager任务使用

10,000,000 文件数量配置

类型	节点数	CPU	内存	硬盘	网络	说明
NameNode	2	8 Cores	≥64 GB	RAID1 SAS ≥600GB	2*10Gb光口	内存和文件对象呈线性比例关系，硬盘主要是持久化元数据和EditLog
JournalNode	3	8 Cores	2 GB	RAID1 SAS ≥600GB	2*10Gb光口	硬盘主要是持久化元数据和EditLog
DataNode	≥30	8 Cores	16 GB	SAS/SATA *N	2*10Gb光口	硬盘用于存储业务数据的blocks，通常是10块以上的4T SATA

说明

容量配置要求主要体现的变更是NameNode的内存和DataNode的节点数。

2.3 NameNode 资源部署

2.3.1 数量配置

容量角度

当前每对NameNode 支持的文件数量不超过1.5亿（小文件），如果业务规划超过，则需要考虑扩容NameNode（支持Federation特性），或者新建集群。或采用小文件处理方案，Hadoop Archive特性或者SmallFS功能。

FusionInsight HD 性能调优指导手册 - HDFS

性能角度

当前默认支持的OPS，全写大概3K，全读在11K左右，可以根据NameNode的RPC的排队时长和处理时长。

2.3.2 资源配置

操作场景

NameNode中文件对象需要占用一定的内存，消耗内存大小随文件对象的生成而线性递增。NameNode中，文件对象可以是文件、目录或者Block。在NameNode WebUI界面的Summary也可以看到文件对象（filesystem objects）的统计。

说明

文件对象计算样例。

- 如果HDFS中有1000个1K的文件，则NameNode需要存储的文件对象为2000（1000文件个数+1000block数量）。
- 如果HDFS中有1000个1G的文件，则NameNode需要存储的文件对象为9000（1000文件个数+1000*（1G÷128M）block数量）。

操作步骤

需要配置NameNode服务端的如下参数：

参数名	描述
GC_OPTS 服务端参数	JVM前4个值为，“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”需要根据文件对象修改为具体的内存设置值参考如下表所示。

说明

对象计算样例NameNode文件数量和NameNode使用的内存大小成比例关系，文件对象变化时请修改默认值中的“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”。参考值如下表所示。注意1亿小文件对应2亿文件对象。根据文件对象设置按比例配置。

部署参考

文件对象数量（filesystem objects=file+block）	参考值
10,000,000	-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
20,000,000	-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G
50,000,000	-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G
100,000,000	-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G
200,000,000	-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G
300,000,000	-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G

监控检查

通过NameNode实例的以下监控项可以监控资源分配是否合理。

- NameNode Memory
- Used Memory of the NameNode
- NameNode Garbage Collection (GC) NameNode

内存的百分比占用率到70% 或者周期内的GC时间超过1s就需要跟踪部署章节的要求调整。

2.4 DataNode 部署

2.4.1 根据容量规划 DataNode 部署个数

指导原则

- 从副本数量角度规划，建议一个DataNode的管理的副本数不要超过500万
  DataNode个数 =（总体blocks*3副本）/5,000,000
- 从存储容量角度规划
  DataNode个数 =（总容量*3副本）/ (DataNode磁盘个数*单个磁盘容量)。

样例说明

- 如果HDFS中有1亿小文件，则默认所有DataNode存储3亿副本，则DataNode需要个数至少建议60左右。

FusionInsight HD 性能调优指导手册 - HDFS

- 如果HDFS规划存储1P数据，默认3副本，总存储容量则需要3P空间。建议存储容量低于80%的情况下。假设每个DataNode 10个4T磁盘，则DataNode，至少需要95个左右。

2.4.2 根据性能吞吐量规划 DataNode 部署个数

指导原则

假设当前每个DataNode 配置10块4T容量的SATA盘，单硬盘写入理论能力在150M/s。

- 当前实际能力按75%左右计算，单DataNode的实际写入能力在1G左右（磁盘和网络之一的瓶颈）。考虑3副本。每个DataNode对外呈现的写入能力在350M/s。
- 每个DataNode的读取能力1G 左右。

样例说明

根据DataNode的吞吐能力规划的节点数计算公式如下：

DataNode个数 = 写HDFS速度要求/350 M+读HDFS要求/1000M

说明

HDFS集群对外写入的吞吐量大概150M/s 每DataNode节点。

2.4.3 DataNode 资源配置

操作场景

DataNode配置的内存资源配置根据管理的副本数成线性关系配置即可。

操作步骤

参数名	描述
GC_OPTS 服务端参数	JVM前4个值为，“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”需要根据副本数修改为具体的内存设置值，具体参考下表。大容量建议配置如下： “-Xms16G -Xmx16G -XX:NewSize=1G -XX:MaxNewSize=2G”

指导原则

每个DataNode副本数量	参考值
2,000,000	-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
4,000,000	-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G

FusionInsight HD 性能调优指导手册 - HDFS

监控检查

通过DataNode实例的以下监控项可以监控资源分配是否合理。

- **Percentage of Used Memory of the DataNode**
- **Percentage of Used Memory of the DataNode**
- DataNode Garbage Collection (GC) DataNode的GC时长

内存的百分比占用率到70% 或者周期内的GC时间超过1s就需要跟踪部署章节的要求调整。

FusionInsight HD 性能调优指导手册 - HDFS

3 典型业务的调优

上层组件业务一般分为可以分为IO密集型业务，计算密集型业务，低延迟业务，高吞吐量业务。

FusionInsight HD 性能调优指导手册 - HDFS

3.1 计算密集型业务

计算密集型业务，低延迟业务，这一类业务通常是NameNode文件目录访问为主的（权限判断，文件存在性，文件信息查询等对NameNode元数据操作为主的）。主要针对NameNode的处理能力进行调整。在NameNode的默认OPS无法支撑的情况下，只能通过调整业务操作对NameNode进行优化或者新增NameNode引入Federation来解决。

3.2 低延迟业务

概述章节已经提过，由于HDFS的架构，这一类业务通常是NameNode文件目录访问为主的（权限判断，文件存在性，文件信息查询等对NameNode元数据操作为主的）。

3.3 高吞吐量业

高吞吐量业务，这一类业务主要是大文件操作，对DataNode的IO需求更高。在默认配置情况下，通过扩容DataNode节点的磁盘和数量。

3.4 IO 密集型业务

IO密集型业务类似于高吞吐量业务 这一类业务对于HDFS类似，都是对DataNode的IO需求更高。在默认配置情况下，通过扩容DataNode节点的磁盘和数量来解决。

3.5 归档业务与实时业务

对于归档数据：

用于存储归档数据，可以使用标签存储，和分级存储，将节点按性能差异划分为不同的类型。

标签存储配置归档目录为慢节点，高密存储节点。

对于实时业务：

用于实时业务数据，可以使用标签存储，和分级存储，将节点按性能差异划分为不同的类型。

标签存储配置归档目录为高性能节点，高密存储节点。

FusionInsight HD 性能调优指导手册 - HDFS

4 二次开发业务应用指导

HDFS的数据导入，读取当前大部分是通过上层组件Flume HBase Loader等实现的

也有部分的HDFS的二次开发针对读写的处理。

应用的客户端应用开发处理通常是采用分布式框架（如MR） 部署在各节点运行。

如果部署在集群外或者部署在单节点上。分布在不同节点的客户端上，避免单节点的客户端网络IO或者磁盘IO成为瓶颈。

针对HDFS的具体应用开发有一些常用的API可以提升处理能力。

- FSDataInputStream.setReadahead (Long readahead)
- FSDataInputStream.setDropBehind (Boolean dropBehind)

FusionInsight HD 性能调优指导手册 - HDFS

5 性能调优常用方法

FusionInsight HD 性能调优指导手册 - HDFS

5.1 优化 NameNode 处理能力

5.1.1 日志处理能力

操作场景

NameNode是所有操作的瓶颈。对于大并发的处理namenode瓶颈的情况下,审计日志的操作对NameNode性能影响比较大，对于大部分集群使用场景，读取的处理远高于写入修改的操作，关闭部分读取的审计日志能提升并发请求的处理能力，关闭日志会影响读写失败的时的问题定位能力。

修改参数

参数名	优化描述
dfs.namenode.audit.log.debug.cmdlist	在hdfs.audit.log.level默认为INFO级别时，可以设置NameNode部分审计日志类型（如读取）不输出，在hdfs.audit.log.level设置为DEBUG时则输出所有审计日志。默认不控制，可以配置以逗号隔开的命令如,**listStatus**,**getfileinfo**,**open**,**getAclStatus**则默认红色的操作记录不会再审计日志中输出。

说明

HDFS审计日志中常见的操作命令如下，蓝色为修改类的操作不建议屏蔽，红色的可以默认屏蔽设置到此配置项中。

listStatus,getfileinfo,open,getAclStatus, isFileClosed, contentSummary

setPermission,setOwner, concat,setTimes,setReplication,truncate,setStoragePolicy，create,append,rename,delete,mkdirs,setQuota，,modifyAclEntries,removeAclEntries,removeAcl,setAcl ,checkAccess

5.1.2 文件完成判断控制

操作场景

通常，一个文件只有在所有Block是complete状态才能够关闭。当这个配置项配置一个正整数值N时，允许文件剩余N个Block是committed状态下，关闭文件。

修改参数

参数名	优化描述
dfs.namenode.file.close.num-committed-allowed	在文件写入最后的一个block还未从DataNode上报状态时，允许客户端关闭文件，从而缩短文件关闭耗时。默认值为0，可以修改为1。

FusionInsight HD 性能调优指导手册 - HDFS

5.1.3 块上报策略控制

操作场景

当前DataNode增量上报处理，每次操作后都会上报，如果设置为一个正整数，则datanode上的incremental blockreport会等待响应的毫秒数时延后，集中上报。

修改参数

参数名	优化描述
dfs.blockreport.incremental.intervalMsec	默认值0，修改值1000。合并增量块上报操作，1000毫秒合并一次，以减少NameNode上因为增量块上报操作导致的频繁锁切换。

5.2 最大线程控制

操作场景

对于大集群场景，需要考虑调整服务的并发处理能力。

修改参数

参数名	优化描述
dfs,namenode.hander.count	设置namenode能够同时处理的请求数，一般为集群规模的自然对数lnN 的20倍，小于100节点的集群设置64就可以，小于1000节点配置128就能满足要求。
dfs.datanode.max.transfer.threads	datanode用于传输数据的最大线程数，默认4096，对于高并发场景调整为8192。
dfs.datanode.handler.count	设置datanode处理rpc的线程数，大集群可以适当加大。

5.3 修改 Client 的设置

说明

HDFS客户端配置文件路径生效在：客户端安装目录/HDFS/hadoop/etc/hadoop /hdfs-site.xml，客户端安装目录/HDFS/hadoop/etc/hadoop /core-site.xml，对于MR等上层引用也可以理解成HDFS的客户端。

FusionInsight HD 性能调优指导手册 - HDFS

5.3.1 修改 Blocksize

操作场景

HDFS作为mapreduce的基础分布式文件系统，对mapred的运行效果也有直接的影响。首先影响到我们的性能的参数。

修改参数

参数名	描述
dfs.blocksize 客户端参数	默认134217728对应128M，针对大文件计算可以设置为256M或者更大。

5.3.2 修改副本数

操作场景

对于大集群场景，需要考虑调整服务的并发处理能力。

修改参数

参数名	描述
dfs.replication 客户端参数	默认3副本。但是由于mapred的map需要输入数据，一般默认情况是一个map一个block，如果有多个map在一个机器执行时，如果这个map的输入数据是本地的，那么显然map的执行将会更快，如果在大集群，对于特定的文件读取可以设置更多的副本数。 类似的使用场景如mapreduce.client.submit.file.replication，可以配置submit阶段的副本数为10。

5.3.3 修改 ipc 设置

操作场景

在对安全要求的不高的场景，可以关闭rpc的加密处理。能提升读写能力。

FusionInsight HD 性能调优指导手册 - HDFS

修改参数

参数名	描述
dfs.client-write-packet-size 客户端参数	当HDFS客户端写数据至数据节点时，数据会被累积，直到形成一个包。然后这个数据包会通过网络传输。dfs.client-write-packet-size配置项可以指定该数据包的大小。这个是客户端参数，可以通过每个job进行指定。 默认值262144，对于高带宽网络修改此值增大可以增加吞吐量。

5.3.4 短路读提高读性能

操作场景

Yarn，HBase ， Impala等业务应用部署在集群内和DataNode共节点的业务。

修改参数

参数名	优化描述
dfs.client.read.shortcircuit 服务端参数	在HDFS中，通常是通过DataNode来读取数据的。但是，当客户端向DataNode请求读取文件时，DataNode就会从磁盘读取该文件并通过TCP socket将数据发送到客户端。所谓“Short Circuit”是允许DataNode从磁盘直接读取文件.默认开启,需要注意,只有客户端也在datanode上共节点共节点才有意义。
dfs.domain.socket.path 客户端和服务端参数要一致	/var/run/FusionInsight-HDFS/dn_socket

5.4 安全设置参数优化

操作场景

在对安全没有特别要求的场景，可以关闭rpc的加密处理。能提升读写能力。

修改参数

参数名	描述
hadoop.rpc.protection 客户端服务端的值要保持一致	当前默认配置privacy，对消息传输做了加密处理，修改为authentication可以提升速度。

FusionInsight HD 性能调优指导手册 - HDFS

5.5 修改 ipc 设置

操作场景

对于大文件，增加DataNode IO的吞吐量进行调整。

修改参数

参数名	描述
io.file.buffer.size	放大HDFS读写的缓存区大小，默认大小为4K，针对好的硬件，可以适当调大，如128K。

5.6 Balancer

操作场景

在Datanode扩容或者异构场景磁盘容量不均时，有时需要做均衡处理，修改以下参数会提升均衡速度。但是占用系统带宽较高。

修改参数

参数名	描述
dfs.datanode.balance.max.concurrent.moves	同时移动文件数，默认为5，可以修改为50。
dfs.namenode.replication.work.multiplier.per.iteration	每一批次移动文件数量，默认为2，可以修改为64。
dfs.datanode.balance.bandwidthPerSec	单节点均衡带宽默认20971520对应20M，需要根据集群实带宽传输能力设置，业务空闲或者带宽能力足够，可以设置为20971520，对应200M。

说明

对于带宽的操作也可以执行以下命令处理，不用重启即可生效。

hdfs dfsadmin -setBalancerBandwidth <bandwidth in bytes per second>

FusionInsight HD 性能调优指导手册 - HDFS

5.7 退服

操作场景

在退服的时候，由于namenode的处理，调度的优先级比写入业务流程的优先级低，因此在退服的速度也会极大的受当前业务的影响，以下的参数优化，在业务量较低的时候有明显的变化。

修改参数

参数名	描述
dfs.namenode.replication.max-streams	副本流处理数量，社区默认值为2，FI默认值修改为64。
dfs.namenode.replication.max-streams-hard-limit	社区默认值4，FI默认值修改为128。

5.8 基于标签的存储策略设置

操作场景

可以通过标签配置，将基于HDFS特定目录的业务调整到部分DataNode进行处理，可以将对优先级高的业务资源进行保证。

修改参数

参数名	描述
dfs.block.replicator.classname	默认为 org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyDefault 修改为 org.apache.hadoop.hdfs.server.blockmanagement.BlockPlacementPolicyWithNodeTag
Host2tag	设置每个节点对应的nodelabel
Path2expression	设置HDFS对应目录和nodelabel间的关系

FusionInsight HD 性能调优指导手册 - HDFS

5.9 基于分级的存储策略设置

操作场景

HDFS的异构分级存储框架提供了RAM_DISK、SSD、DISK、ARCHIVE四种存储类型的存储设备，以对应DataNode上可能存在的不同的存储介质。

- RAM_DISK是一种由内存虚拟的硬盘，具有最高的读写性能。其容量受限于内存大小，通常容量很小，且掉电可能丢失数据。
- SSD即固态硬盘，具有较高的读写性能。但通常存储容量较小，单位存储成本比普通机械硬盘高。
- DISK即普通机械硬盘，是HDFS用于保存数据的主力存储类型。
- ARCHIVE类型代表高密度低成本的存储介质，读写性能相对较差，通常装配于计算能力较低的节点，用于大容量非热点数据存储。

修改参数

参数名	描述
dfs.datanode.data.dir	默认情况下系统认为数据保存的存储设备为DISK，此时需要根据实际存储设备的类型修改，参数值为“[存储设置类型]存储目录”，多个目录使用逗号隔开。修改效果如下：“[RAM_DISK]/home/hadoop/dfs/ram,[SSD]/home/hadoop/dfs/ssd,/home/hadoop/dfs/hd,[ARCHIVE]/home/hadoop/dfs/archive”。

5.10 清除缓存

操作场景

在有些场景下，特别是对一些大的，并且不可能重用的数据，缓存在操作系统的缓存区是无用的。此时，DN可以配置自动清理缓存区数据，在已经发生向客户端之后。此功能自动失效，在读取小数据片时。(例如HBase的随机读写场景）。通过释放缓存，这在某些场景下可以提高性能。

修改参数

参数名	描述
dfs.datanode.drop.cache.behind.reads	默认true
dfs.datanode.drop.cache.behind.writes	默认true

FusionInsight HD 性能调优指导手册 - HDFS

5.11 IO 隔离

操作场景

HDFS DataNode I/O隔离特性的主要功能是可以为每个HDFS用户通过DataNode产生的磁盘I/O资源带宽提供隔离保证，即支持为每个用户配置I/O权重（weight），当多个用户在共享同一磁盘I/O时将按照预设权重分享磁盘的I/O资源。

修改参数

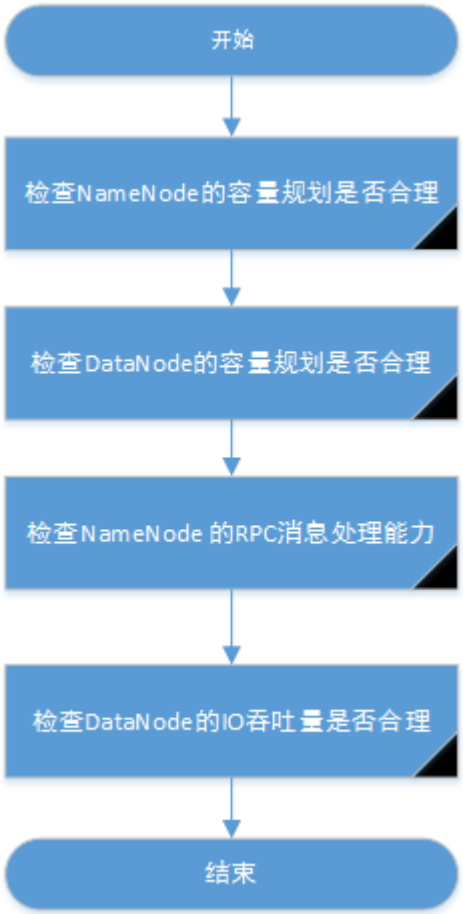
参数名	描述
dfs.datanode.disk-io-tuner.enabled	默认false，关闭打开/关闭隔离特性。
iostat.enabled	决定是否开启iostat进程来监控节点的I/O负载，开启此配置是开启I/O隔离特性（使用DefaultDiskIOEvaluator）和Load Balance特性的先决条件。默认值false表示关闭该功能。
dfs.client.io-weight	输入输出流的默认I/O权重值。比重的取值范围为[1-100]的整数，当用户设置小于1的整数时，默认为1；当用户设置大于100的整数，默认为100；其他异常设置均为默认权重。默认值为10，对于优先级高的组件，根据业务情况调整此值10~100之间。值越高，相对优先级更高。
dfs.datanode.io-settings	高级配置。根据IO隔离特性的配置原则修改。可以针对用户级别的权重进行控制。输入为一个xml文件。

FusionInsight HD 性能调优指导手册 - HDFS

6 性能瓶颈监控及调优

FusionInsight HD 性能调优指导手册 - HDFS

6.1 常见性能问题及解决方案



说明

HDFS的常见问题通常是NameNode的RPC处理能力不足，造成操作的延时较长，或者DataNode的IO能力不足导致读写的吞吐量不够。

6.1.1 NameNode 容量问题

问题描述

NameNode中文件对象需要占用一定的内存，消耗内存大小随文件对象的生成而线性递增。NameNode中，文件对象可以是文件、目录或者Block。在NameNode WebUI界面的Summary也可以看到文件系统对象（filesystem objects）的统计。

从Manager访问“Services > HDFS status > NameNode(Active) ”的Summary 页面查看HDFS的“Files and Directories ”以及“Block”的情况，根据NameNode资源配置章节的总对象和NameNode内存配置值是否一致。

当前每对NameNode 支持的文件数量不超过1.5亿（小文件），如果业务规划超过，则需要考虑扩容NameNode（支持Federation特性），或者新建集群。或采用小文件处理方案，Hadoop Archive特性或者SmallFS功能。

FusionInsight HD 性能调优指导手册 - HDFS

说明

文件系统对象为files directory 和blocks数量之和,计算样例如下

- 如果HDFS中有1000个1K的文件，则NameNode需要存储的文件对象为2000（1000文件个数+1000block数量）
- 如果HDFS中有1000个1G的文件，则NameNode需要存储的文件对象为9000（1000文件个数+1000*（1G÷128M）block数量）。

监控检查

- 通过NameNode实例的以下监控项可以监控当前容量下的资源分配是否合理

NameNode Memory

Used Memory of the NameNode

NameNode Garbage Collection (GC) NameNode

内存的百分比占用率到70% 或者周期内的GC时间超过1s就需要跟踪部署章节的要求调整

- 通过ALM-14006 HDFS文件数超过阈值告警，可以设置集群的预警阈值。

当前告警默认阈值10,000,000。根据设置参考部分的配置进行调节，可以做到提前预警。

说明

对象计算样例NameNode文件数量和NameNode使用的内存大小成比例关系，文件对象变化时请修改默认值中的“-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M”。参考值如下表所示。注意1亿小文件对应2亿文件系统对象。根据文件对象设置按比例配置

操作修改

需要配置NameNode服务端的如下参数

FusionInsight HD 性能调优指导手册 - HDFS

参数名	描述
NameNode实例的GC_OPTS 服务端参数	JVM前4个值为，-Xms2G -Xmx4G -XX:NewSize=128M -XX:MaxNewSize=256M需要根据文件对象修改为具体的内存设置值参考如下

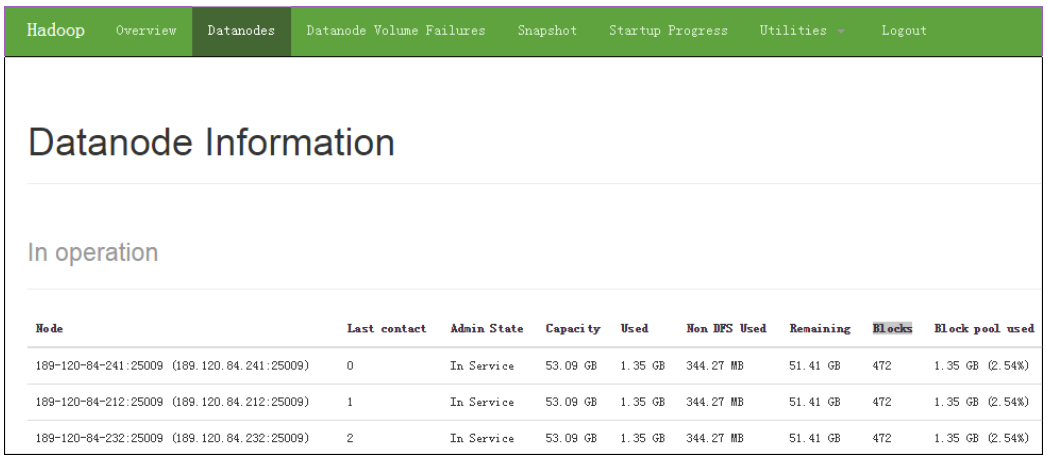
设置参考

文件对象数量（filesystem objects=file+block）	参考值
10,000,000	-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
20,000,000	-Xms12G -Xmx12G -XX:NewSize=1G -XX:MaxNewSize=1G
50,000,000	-Xms32G -Xmx32G -XX:NewSize=3G -XX:MaxNewSize=3G
100,000,000	-Xms64G -Xmx64G -XX:NewSize=6G -XX:MaxNewSize=6G
200,000,000	-Xms96G -Xmx96G -XX:NewSize=9G -XX:MaxNewSize=9G
300,000,000	-Xms164G -Xmx164G -XX:NewSize=12G -XX:MaxNewSize=12G

6.1.2 DataNode 容量问题

问题描述

从Manager访问“Services > HDFS status > NameNode(Active) ”页面查看每个DataNodes的Block数量和磁盘使用量，需要根据前面资源配置方式考虑调整DataNode的内存。



监控检查

通过DataNode实例的以下监控项可以监控资源分配是否合理

- **Percentage of Used Memory of the DataNode**
- **Percentage of Used Memory of the DataNode**
- DataNode Garbage Collection (GC) DataNode的GC时长

内存的百分比占用率到70% 或者周期内的GC时间超过1s就需要跟踪指导原则的章节要求调整

部署原则

- 从副本数量角度规划，建议一个DataNode的管理的副本数不要超过500万
  DataNode个数 =（总体blocks*3副本）/5,000,000
- 从存储容量角度规划
  DataNode个数 =（总容量*3副本）/ (DataNode磁盘个数*单个磁盘容量)。

样例说明

- 如果HDFS中有1亿小文件，则默认所有DataNode存储3亿副本，则DataNode需要个数至少建议60左右。
- 如果HDFS规划存储1P数据，默认3副本，总存储容量则需要3P空间。建议存储容量低于80%的情况下。假设每个DataNode 10个4T磁盘，则DataNode，至少需要95个左右。

设置参考

每个DataNode副本数量	参考值
2,000,000	-Xms6G -Xmx6G -XX:NewSize=512M -XX:MaxNewSize=512M
5,000,000	-Xms16G -Xmx16G -XX:NewSize=1G -XX:MaxNewSize=2G

6.1.3 NameNode 的 RPC 处理能力

NameNode 的 RPC 处理能力

- **Average RPC Queuing Time** rpc在交互中平均等待时间体现的是消息排队时长，时间越长表明请求的消息数量
- **Average RPC Processing Time** 体现的是NameNode对RPC 消息的平均时长

以上两个监控统计值的单位是ms。现场一般小于10ms。如果到百ms级别就比较慢了，表明需要调整NameNode的处理能力（内存，修改日志，引入Federation等方式，参考后续描述）。

- **Used CPU Cores of NameNode NameNode** 使用CPU的情况，过大需要调整分配的CPUset值

- **NameNode Garbage Collection (GC)** NameNode的GC时长

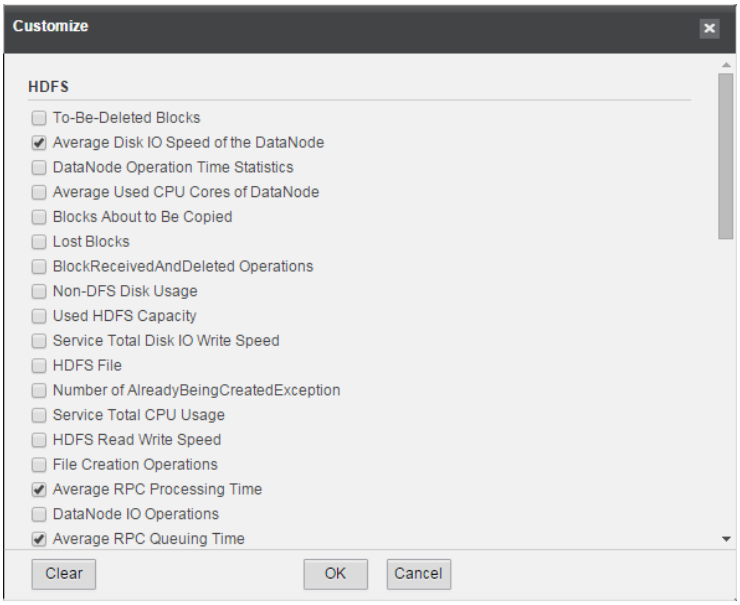
说明

NameNode的处理能力可以通过调整日志处理等一些特性实现。或者通过启用Federation特性新增NameNode实现。

- **Used CPU Cores of NameNode** NameNode使用CPU的情况，过大需要调整分配的CPUset值。

  NameNode的响应情况可以在关注HDFS Service 的监控指标。

  在HDFS服务的监控页面主要关注以下几个监控项：



6.1.4 DataNode 的 IO 处理能力

DataNode 的 IO 处理能力

- **DataNode Packet Operation Time** DataNode的每个数据包的操作时长。每个数据包默认64K，如果时长超过1s，则DataNode的处理IO能力不足。

检查对应节点的网络带宽和磁盘IO的检测，判断是否达到系统硬件资源的限制。

是否触发ALM-12033 慢盘故障。

每一秒执行一次对应的iostat命令，监控磁盘I/O的系统指标，读写和svctm值等参数。

- **Disk IO Speed of the DataNode** DataNode的IO平均读取写入时长。

说明

对于DataNode的处理能力不足情况的调整，如果集群中DataNode实例IO能力已经到达最大值，最直接的方式是增加DataNode挂载磁盘的数量，或者扩容DataNode。

6.1.5 网络瓶颈

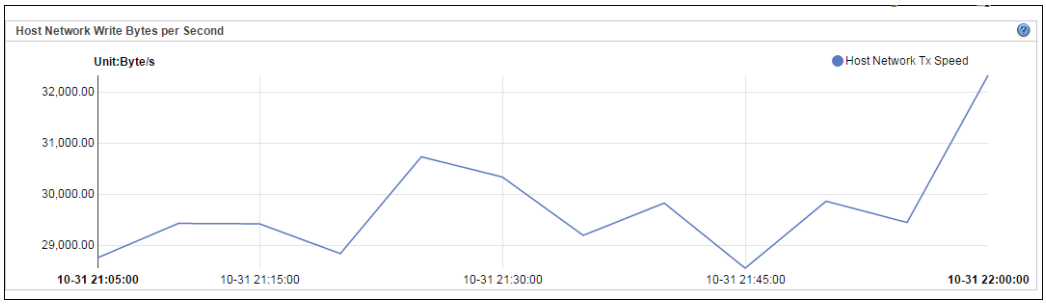
网络瓶颈

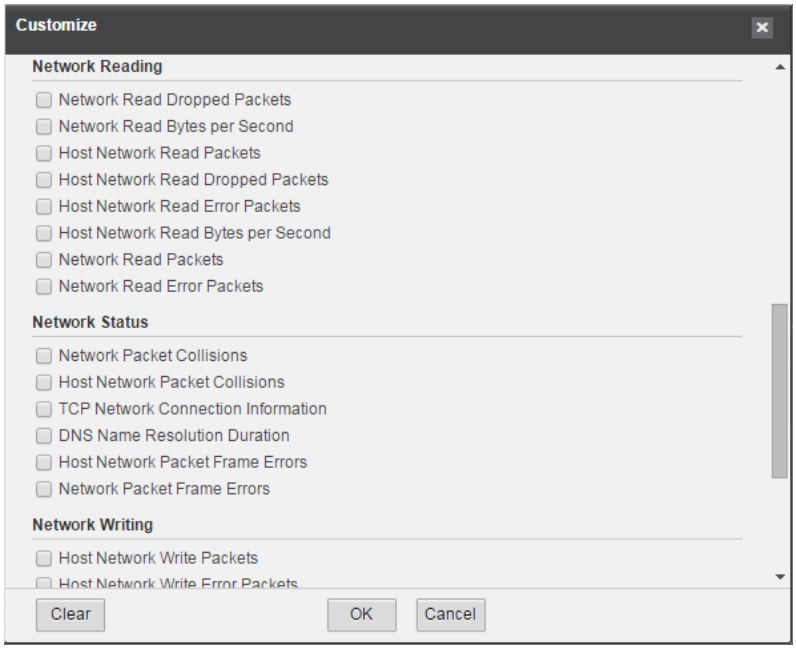
节点写入监控

FusionInsight HD 性能调优指导手册 - HDFS

查询节点的网络监控是否达到网卡的吞吐能力。也需要关注节点对应的汇聚交换机的处理能力是否通过以下监控节点网络的处理情况。

Host Network Write Bytes per Second





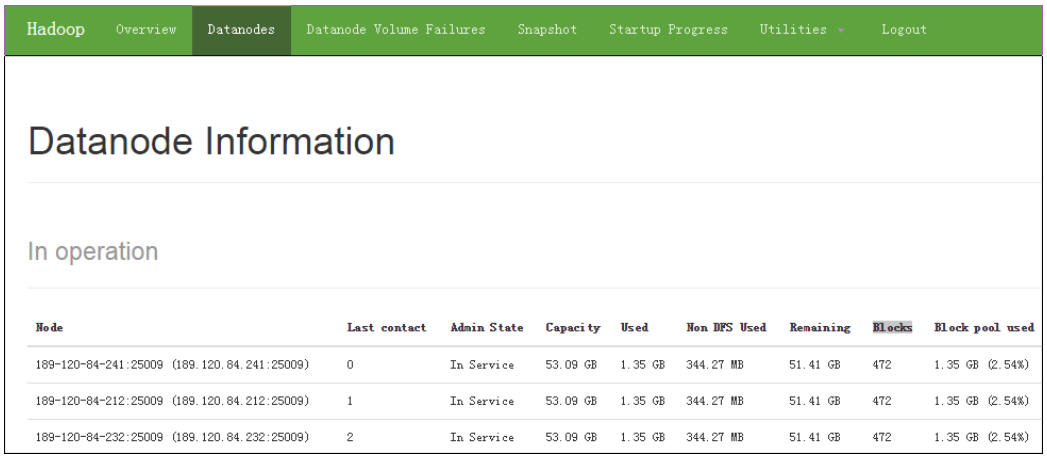
6.2 监控手段

6.2.1 容量监控

从Manager访问“Services > HDFS status > NameNode(Active) ”的Summary 页面查看HDFS的“Files and Directories ”以及“Block”的情况，根据NameNode资源配置章节的总对象和NameNode内存配置值是否一致。

FusionInsight HD 性能调优指导手册 - HDFS

从Manager访问“Services > HDFS status > NameNode(Active)”页面查看每个DataNodes的Block数量和磁盘使用量，需要根据前面资源配置方式考虑调整DataNode的内存。



当前每对NameNode 支持的文件数量不超过1.5亿（小文件），如果业务规划超过，则需要考虑扩容NameNode（支持Federation特性），或者新建集群。或采用小文件处理方案，Hadoop Archive特性或者SmallFS功能。

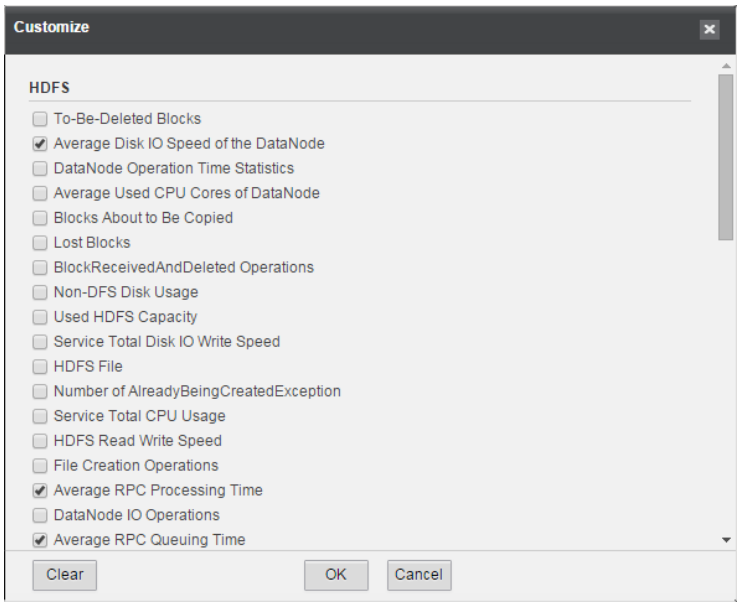
6.2.2 性能监控

HDFS 服务监控

NameNode的响应情况可以在关注HDFS Service 的监控指标。

在HDFS服务的监控页面主要关注以下几个监控项：

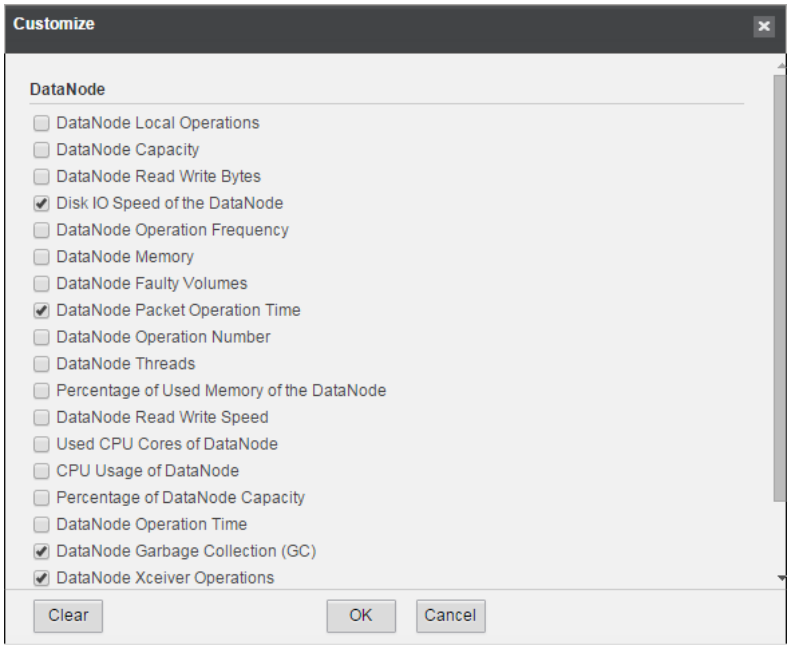
FusionInsight HD 性能调优指导手册 - HDFS



- **Average RPC Queuing Time** rpc在交互中平均等待时间体现的是消息排队时长，时间越长表明请求的消息数量
- **Average RPC Processing Time** 体现的是NameNode 对RPC 消息的平均时长

以上两个监控统计值的单位是ms。现场一般小于20ms。如果到秒级别就比较慢了，表明需要调整NameNode的处理能力（内存，修改日志，引入Federation等方式，参考后续描述）

DataNode 实例监控



- **Disk IO Speed of the DataNode** DataNode的IO平均读取写入速度。
- **DataNode Packet Operation Time** DataNode的每个数据包的操作时长。每个数据包默认64K，如果时长超过1s，则DataNode的处理能力不足。

FusionInsight HD 性能调优指导手册 - HDFS

- **DataNode Xceiver Operations** DataNode的并发的流处理数量，当前默认最大值8192，如果过大需要调整。

FusionInsight HD 性能调优指导手册 - HDFS

7 POC 调优指导

FusionInsight HD 性能调优指导手册 - HDFS

7.1 常见 HDFS 标准测试场景样例

操作场景

此外除了只调整HDFS参数外，涉及到读写吞吐量的业务处理除了与服务端的参数设置外，或者HDFS客户端参数，也与业务的并发量相关，以下针对几种使用场景，提供样例作为参考。

操作步骤

用例	描述
TestDFSIO read	hadoop jar hadoop-mapreduce-client-jobclient-*-tests.jar TestDFSIO \ -Dmapreduce.reduce.merge.memtomem.enabled=true \ -Dmapreduce.job.reduce.slowstart.completedmaps=0.0f \ -Dmapreduce.map.cpu.vcores=10 \ -Ddfs.datanode.max.transfer.threads=8192 \ -Dmapreduce.map.memory.mb=4096 \ -Dmapreduce.reduce.memory.mb=4096 \ -Dmapreduce.reduce.java.opts=-Xmx3584M \ -Dmapreduce.map.java.opts=-Xmx3584M \ -Dmapred.child.java.opts=-Xmx3584M \ -Dio.file.buffer.size=1048576 \ -read -nrFiles 100 -fileSize 100000
TestDFSIO write	hadoop jar hadoop-mapreduce-client-jobclient-*-tests.jar TestDFSIO \ -Dmapreduce.job.reduce.slowstart.completedmaps=1 \ -Dmapreduce.map.cpu.vcores=10 \ -Ddfs.datanode.max.transfer.threads=8192 \ -Dmapreduce.map.memory.mb=4096 \ -Dmapreduce.reduce.memory.mb=4096 \ -Dmapreduce.reduce.java.opts=-Xmx3584M \ -Dmapreduce.map.java.opts=-Xmx3584M \ -Dmapred.child.java.opts=-Xmx3584M \ -Dmapreduce.reduce.shuffle.parallelcopies=10 \ -Dmapreduce.task.io.sort.mb=1024 \ -write -nrFiles 100 -fileSize 100000

用例	描述
nnbench	hadoop jar ./hadoop-test.jar nnbench -operation create_write \ -maps 80 -reduces 40 -bytesToWrite 1 -numberOfFiles 10000 -blockSize 1048576 \ -replicationFactorPerFile 3 -readFileAfterOpen true -baseDir /benchmarks/NNBench
Distcp	根据资源合理指定map个数 hadoop distcp hdfs://nn1:25000/source/first hdfs://hacluster/target -m 100

7.2 特定场景提升方法

7.2.1 Centralized Cache 提高读性能

操作场景

在Datanode节点有大量空余内存,并且对读响应有严格要求时可以使用。由于使用此特性占用大量内存，因此在有大量剩余内存或者没有Yarn的 Container使用内存的场景，如（ELK部署）。属于高级使用能力，慎用。

修改参数

参数名	描述
/etc/security/limits.conf	FI现在版本尚未自动化适配内存cache配置,需要手动在每个DataNode节点上修改并重启节点生效。
dfs.datanode.max.locked.memory	DataNode锁定内存，将特定目录的HDFS文件内容同时缓存到到内存池中，比磁盘读取快。

说明

shell具体使用说明参考样例**https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-hdfs/centralizedcachemanagement.html**

7.2.2 写同步机制和 cache

操作场景

如果没有电源保护，DataNode异常掉电时，系统缓存中接收到的Client端写的还未来得及刷到磁盘的数据会丢失。由于安全性要求，保证DataNode异常掉电时，尽可能地减少数据丢失的量，DataNode会定期sync系统缓存中的数据到磁盘。但sync到磁盘会降低写的效率，降低集群的整体吞吐量。

FusionInsight HD 性能调优指导手册 - HDFS

修改参数

参数名	描述
dfs.datanode.sync.behind.writes	DataNode是否指示操作系统在写数据后立即将这些数据写入磁盘。

如果系统有完善的电源保护系统且对少量数据丢失不明感，可以将该值修改为false 以能提升写入能力。对于异常掉电。可能会造成数据丢失。

7.2.3 通过减少客户端的副本数设置

操作场景

对于写入有严格要求的POC测试，可以通过减小副本数降低可靠性的方式提升写入的性能。

操作步骤

修改参数

参数名	描述
dfs.replication 客户端参数	默认3副本。但是可以修改客户端的副本数设置，修改客户端对应的副本数设置。

7.2.4 磁盘 cache

操作场景

开启磁盘cache提升写入能力。

限制

机器意外断电可能会造成信息丢失

操作步骤

- **开启磁盘cache**

sdparm --set=WCE /dev/sdx 开启对应磁盘的Disk Write Cache

- **关闭磁盘cache**

sdparm --clear=WCE /dev/sdx 关闭对应磁盘的Disk Write Cache