分布式文件系统 HsDFS 使用手册

目录

	联系人	电话	Email	时间	版本	备注
编写	周超勇	13910123864	bgnvendor@gmail.com	2013.08.24	v0.1	初稿
审核						

1 引言

1.1 编写目的

本手册介绍分布式文件系统 HsDFS 的适用范围、安装配置和操作使用等。

1.2 适用范围

1.2.1 目标用户适用范围

(1) 分布式文件系统 HsDFS 个人研发者
(2) 分布式文件系统 HsDFS 单位研发者
(3) 分布式文件系统 HsDFS 运营维护者

1.2.2 硬件环境适用范围

(1) 单台服务器最低硬件配置要求：500MHZ CPU 主频，800M 内存，1G 硬盘，1 个千兆网卡
(2) 集群最低硬件配置要求：1 个物理节点，1 个千兆以太网交换机（可选），500MHZ CPU 主频，1G 内存，2G 硬盘，千兆网卡
(3) 集群规模至多 xx 台，网络带宽 1000Mbps 及以上

1.2.3 软件环境适用范围

(1) 服务器使用操作系统：Centos 5.x, 6.x X86_64，其它 Linux 系统请测试后确认。
(2) 服务器操作系统最低版本要求：Linux 内核 2.6.32 版本
(3) 集群环境要求不能混搭 64 位与 32 位操作系统

1.2.4 发布形式与许可证

分布式文件系统 HsDFS 以源码压缩包的形式发布，并遵循分布式计算平台相同的许可证：

平台采用 GPLv2 开源许可证发布。按照 GPLv2 开源许可证，其内容包括：使用了 GPLv2 代码的软件在发布时必须使用 GPLv2 许可证发布源代码。

为了便于推广使用，平台许可证设置了一些例外，即在 GPLv2 许可证之外可以向使用者发布商业应用的商业许可证。目前平台全部代码均为作者周超勇编写。

当用户遵循以下两个前提条件时：

1. 在平台中保留 LOGO
2. 在产品说明中声明使用本平台

经审核后发放相应的商业许可证。这个申请流程是：

1、用户提出申请，并填写下面的申请表格给出相应的一些信息

2、审核后给出商业许可证协议的电子文本

3、用户对电子文本无异议后，打印商业许可协议并盖章，然后快递给我们

4、我们收到后，盖章或签字后再寄回给企业。

这其中寄送的快递费用由用户承担。免费商业许可证申请表格如下：

公司名称	
公司地址	
联 系 人	
联系电话	
申请条件	☐ 保留分布式计算平台 LOGO ☐ 在产品使用说明书中声明使用分布式计算平台
产品简介	
（包含：产品简要介绍、应用领域、使用到的分布式计算平台功能情况）	
产品图片	
（申请免费商业许可证，此项是必须的）	

表格信息填写完毕后，需要邮件发送到：bgnvendor@gmail.com。通常我们收到申请后会给出回复，如果没有收到回复，有可能被 Google Mail 当作垃圾邮件处理了。遇到这种情况，请多发送几次。

对于不满足前面提到的条件的企业用户，或希望申请纯商业许可证的企业用户（例如考虑到商业保密，技术保障支持服务等），也可以向我们购买纯商业使用授权【注 2】。

【注 1】

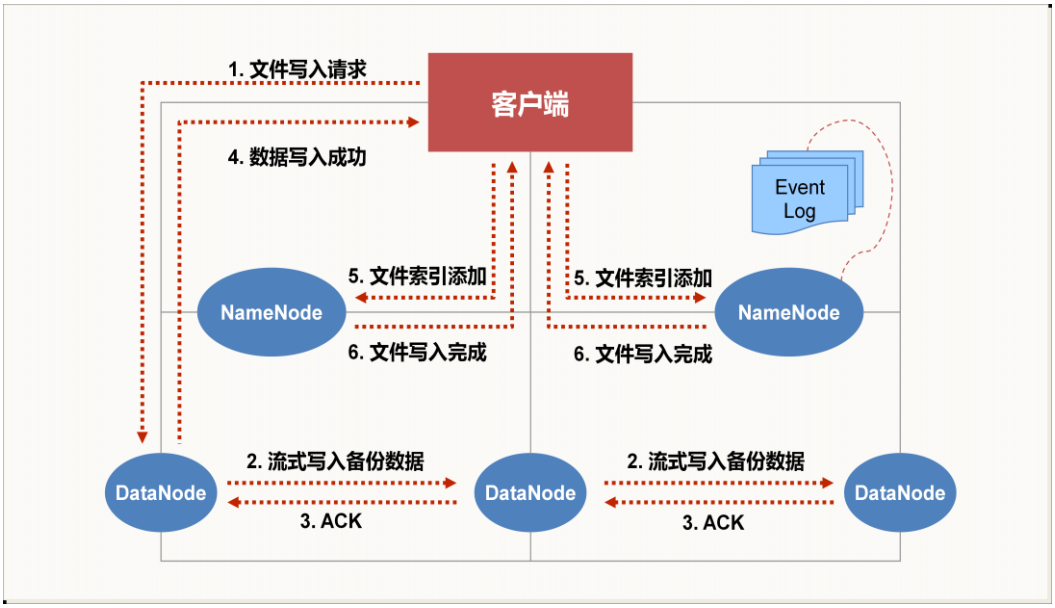
- 对于企业已经开始销售的产品，我们有权不发放商业许可证，而是采用 GPLv2 许可证处理；
- 对于未有商业许可证且未按照 GPLv2 许可证发布代码的用户，我们将联合国内开源届的基金协会、律师向企业提起诉讼以维护我们的权利；

【注 2】可以向 bgnvendor@gmail.com 联系纯商业使用授权，并在邮件文本上进行注明。

2 介绍

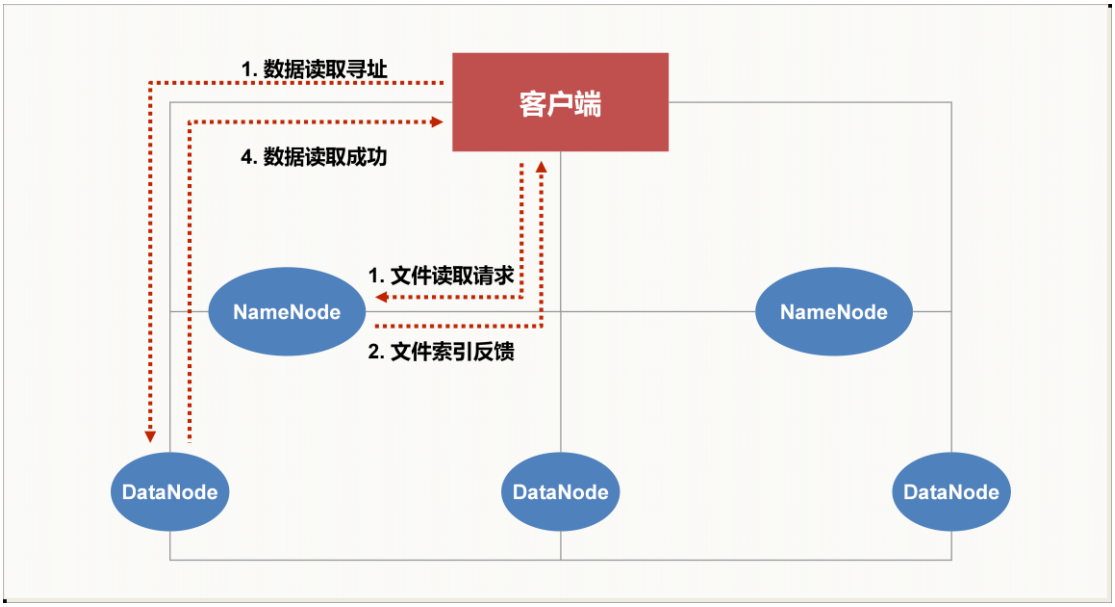
基于分布式计算平台 BGN 的针对海量小文件(文件尺寸 < 64MB)的分布式文件系统(HSDFS)，满足当前电商、互联网分析与搜索、交通、科研等需求，支持高并发，支持 PB 级存储量，在普通 7200 转 SATA 机械磁盘上日吞吐量超过 4TB。

HsDFS 采用类似 Hadoop 的 NameNode + DataNode 组网方式设计。其中，Name Node 储存文件位置信息和文件目录结构信息，采用内存缓存热点信息，支持百亿级文件量；Data Node 存储文件内容，多个小文件组成一个 64M 数据块，Data Node 内存缓存热点数据块。每个小文件支持 3 个备份。



图例说明：

根据实时负载状况，文件内容流式写入各个 Data Node，文件索引信息同时写入互为备份的两个 Name Node 中。仅部分成功的流程将触发日志写入操作，以备后期完备化。



图例说明：

根据实时负载状况，从负载较轻的 Name Node 提取文件索引信息，根据索引信息和实时负载状况，从负载较轻的 Data Node 读取文件内容。

3 安装配置

3.1 编译前准备

(1) 从安装盘的 RPM 库中安装并确认 GCC，GDB
(2) 从安装盘的 RPM 库中安装并确认 LIBXML2
(3) 从安装盘的 RPM 库中安装并确认 LIBPCRE

3.2 编译项设置

编译项在 tbd.mk 文件的 CMACRO 中定义，重点关注如下编译项

编译项	缺省值	说明
CTHREAD_STACK_MAX_SIZE	64KB	线程函数栈大小
CTHREAD_STACK_GUARD_SIZE	4KB	线程函数栈保护段大小
CSOCKET_SOSNDBUFF_SIZE	64KB	socket 发送缓冲大小
CSOCKET_SORCVBUFF_SIZE	64KB	socket 接收缓冲大小
CSOCKET_HEARTBEAT_INTVL_NSEC	10 seconds	TCP 连接心跳检测间隔时长
TASK_DEFAULT_LIVE_NSEC	120 seconds	任务缺省超时时长
ROUTINE_SUPPORT_CTHREAD_SWITCH 【*】	on	支持线程池

CROUTINE_SUPPORT_COROUTINE_SWITCH【*】	off	支持协程池
CFUSE_SUPPORT_SWITCH	on	是否支持 FUSE

【*】线程池或协程池只能二选一。

3.3 编译

执行编译命令：

```
make tbd tbd.mk
```

执行清除命令：

```
make clean tbd.mk
```

3.4 安装

分布式文件系统 HsDFS 的安装仅需要将编译所得的二进制可执行文件拷贝到目标服务器的目录下即可，具体目录无限定；在目标服务器下，执行命令，

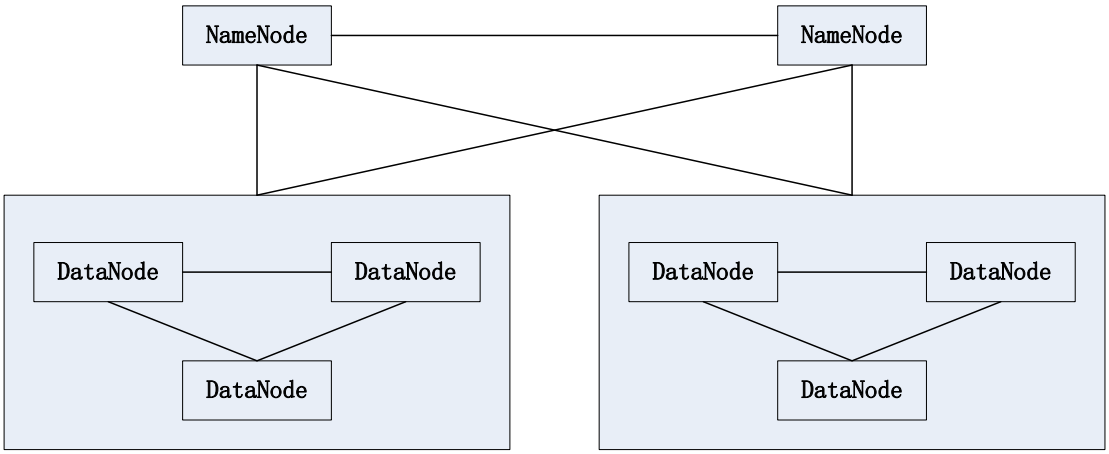
```
ldd tbd
```

检查依赖的库文件是否有缺失。如果存在缺失，请安装相应的软件包，或者从其它相应的服务器上拷贝库文件到缺失路径下。

✧ **警告：请勿将编译后的二进制文件或库文件拷贝到不同类型或版本的操作系统上。**

3.5 配置

分布式文件系统 HsDFS 的配置可参考《分布式计算平台 BGN》的“平台配置文件说明”部分。这里进一步以 2 x NameNode + 3 x DataNode 的组网模型解释配置文件 config.xml。



两个 NameNode 节点全连，每三个 DataNode 一组全连，NameNode 与 DataNode 全连，不同的 DataNode 组之间无需连接。

集群配置部分：

```
<cluster id="1" name="hsdfs_01" model="hsdfs">
    <node role="namenode" tcid="10.10.10.1" rank="0" npdir="/home/ezhocha/hsdfs/10.10.10.1"/>
    <node role="namenode" tcid="10.10.20.1" rank="0" npdir="/home/ezhocha/hsdfs/10.10.20.1"/>
    <node role="datanode" tcid="10.10.10.2" rank="0" dndir="/home/ezhocha/hsdfs/10.10.10.2" group="dn_grp_01"/>
    <node role="datanode" tcid="10.10.10.3" rank="0" dndir="/home/ezhocha/hsdfs/10.10.10.3" group="dn_grp_01"/>
    <node role="datanode" tcid="10.10.10.4" rank="0" dndir="/home/ezhocha/hsdfs/10.10.10.4" group="dn_grp_01"/>
    <node role="client" tcid="10.10.10.5" rank="0"/>
    <node role="client" tcid="10.10.10.6" rank="0"/>
    <node role="client" tcid="10.10.10.7" rank="0"/>
    <node role="client" tcid="10.10.10.8" rank="0"/>
    <node role="client" tcid="10.10.10.91" rank="0"/>
    <node role="client" tcid="10.10.10.92" rank="0"/>
    <node role="client" tcid="10.10.10.101" rank="0"/>
    <node role="client" tcid="10.10.10.102" rank="0"/>
    <node role="client" tcid="10.10.10.201" rank="0"/>
    <node role="client" tcid="10.10.10.202" rank="0"/>
    <node role="client" tcid="10.10.30.1" rank="0"/>
    <node role="client" tcid="10.10.30.2" rank="0"/>
    <node role="client" tcid="10.10.30.3" rank="0"/>
    <node role="client" tcid="10.10.30.4" rank="0"/>
    <node role="client" tcid="10.10.30.5" rank="0"/>
    <node role="client" tcid="10.10.30.7" rank="0"/>
</cluster>
```

分布式文件系统 HsDFS 每个任务通信子只有 1 个进程，即都部署在 0 号进程上（rank = 0），组网模型为“hsdfs”

NameNode: 10.10.10.1， 10.10.20.1

DataNode：每三个一组，其中组“dn_grp_01”的有 10.10.10.1，10.10.10.2，10.10.10.3。

Client：其余为客户端，可部署 HsDFS 的读、写等客户端，支持并发。

4 使用指南

4.1 启动节点的命令

按 NameNode、DataNode、Client 顺序依次启动

(1) 启动 NameNode

在 10.10.10.1 所部署的物理节点上，执行命令

```
./tbd -tcid 10.10.10.1
```

在 10.10.20.1 所部署的物理节点上，执行命令

```
./tbd -tcid 10.10.20.1
```

(2) 启动 DataNode

在 10.10.10.2 所部署的物理节点上，执行命令

```
./tbd -tcid 10.10.10.2
```

在 10.10.10.3 所部署的物理节点上，执行命令

```
./tbd -tcid 10.10.10.3
```

在 10.10.10.4 所部署的物理节点上，执行命令

```
./tbd -tcid 10.10.10.4
```

(3) 启动调测口

比如，在 0.0.0.64 所部署的物理节点上，执行命令

```
./tbd -tcid 0.0.0.64
```

(4) 启动 Client

比如，在 10.10.10.5 所部署的物理节点上，执行命令

```
./tbd -tcid 10.10.10.5
```

4.2 停止节点的命令

比如，停止节点 10.10.10.1，可启动调测口后，在 console 口输入

```
bgn>shutdown work tcid 10.10.10.1
```

停止调测口 0.0.0.64 的命令则是

```
bgn>shutdown dbg tcid 0.0.0.64
```

4.3 创建 HsDFS 流程

(1) 依次启动 NameNode、DataNode 和调测口

(2) 通过调测口创建 NameNode

```
hsdfs create npp /home/hansoul/hsdfs/10.10.10.1 mode 4G max 2 disk max 4 np on tcid 10.10.10.1 at console
```

表示在节点 10.10.10.1 上创建 NameNode。每个 NameNode 块大小为 4GB，最多 2 个 NameNode 块，4 块磁盘。NameNode 持久化数据存放在目录

```
/home/hansoul/hsdfs/10.10.10.1
```

4 块磁盘的路径为

```
/home/hansoul/hsdfs/10.10.10.1/dsk0
/home/hansoul/hsdfs/10.10.10.1/dsk1
/home/hansoul/hsdfs/10.10.10.1/dsk2
/home/hansoul/hsdfs/10.10.10.1/dsk3
```

创建 NameNode 10.10.20.1 的方式与上面类似，命令为

```
hsdfs create npp /home/hansoul/hsdfs/10.10.20.1 mode 4G max 2 disk max 4 np on tcid 10.10.20.1 at console
```

(3) 通过调测口创建 DataNode

```
hsdfs create dn /home/ezhocha/hsdfs/10.10.10.2 with 4 disk 400 GB on tcid 10.10.10.2 at console
```

表示在节点 10.10.10.2 上创建一个 DataNode，其拥有 4 块磁盘，占用每块磁盘 400GB 空间存放数据。DataNode 持久化数据存放在目录

```
/home/hansoul/hsdfs/10.10.10.2
```

4 块磁盘的路径为

```
/home/hansoul/hsdfs/10.10.10.2/dsk0
/home/hansoul/hsdfs/10.10.10.2/dsk1
/home/hansoul/hsdfs/10.10.10.2/dsk2
/home/hansoul/hsdfs/10.10.10.2/dsk3
```

创建其余 DataNode 的方式与上面类似，命令为

```
hsdfs create dn /home/ezhocha/hsdfs/10.10.10.3 with 100 disk 400 GB on tcid 10.10.10.3 at console
hsdfs create dn /home/ezhocha/hsdfs/10.10.10.4 with 100 disk 400 GB on tcid 10.10.10.4 at console
```

(4) 依次停止 DataNode、NameNode
(5) 依次重新启动 NameNode、DataNode

4.4 使用 HsDFS

HsDFS 创建完成后即可使用，操作范例见入口函数 main_cdfs。支持高并发，可开启多个客户端。

5 问题反馈

请保留日志文件，如果有 core dump 文件最好，详细描述软硬件环境、问题的触发条件、观察到的现象等信息，发送至 bgnvendor@gmail.com

6 常见问题