BIG DATA

7.大数据存储

7.1 大数据存储概述

- 大数据存储模型
  - 无格式的文件数据存储
    - byte[]
  - 有格式的文件数据存储
    - block

需要考虑的问题

- 容量
  - PB级
- 延迟
  - 高速缓存、固态存储
- 安全
  - 数据计算，交叉访问权限
- 成本
  - 减少存储消耗、提高复用率、简化管理手段
- 长期保存
  - 对数据定期检测，保证数据长期可用性
- 灵活性
  - 存储系统能够灵活扩展，避免数据迁移

亚马逊发现用卡车转移数据比网络更快

云计算巨头亚马逊称，利用卡车将大型公司客户数据中心的数据转移至其公有云计算设施比使用网络传输更快。亚马逊在年度客户大会上展示了一辆大卡车，这辆大卡车拖着一个名为Snowmobile的庞大存储设备。Snowmobile是一个45英尺长的船运集装箱，数据容量为100PB。亚马逊计划将Snowmobile开到客户办公室提取数据，然后再开往该公司的一处设施，将数据传输到云计算网络，所需时间要远少于通过网络传输。

亚马逊计划将Snowmobile开到客户办公室提取数据

根据亚马逊的计算，10个Snowmobile把从客户办公室存储设备提取的1EB数据传输到亚马逊云端所需时间将降至略低于六个月，而利用高速互联网传输则需26年左右。

7.2 存储方式

- 直接连接存储DAS
- 网络连接存储
- 存储域网络存储
- IP-SAN

直接连接存储

- 直接连接存储是指使用线缆将存储设备直接连接到主机上。
- 此时文件和数据管理依赖本机操作系统。
- 优点：
  - 中间环节少
  - 磁盘读写率高
  - 成本低
- 缺点：
  - 扩展能力有限
  - 数据存储占用主机资源，使得主机性能受到影响
  - 主机故障容易传播为存储故障

直接连接存储分类

- ATA
- SATA
- FC
- SCSI
  - 经常采用的
  - 与服务器连接距离不超过10米
  - 可连接服务器数量有限
  - 受固化控制器限制，无法扩展

直接连接存储的环境

1)小型网络

因为小型网络的规模较小，数据存储量小，而且也不是很复杂，采用直接连接存储方式对服务器的影响不会很大，并且这种存储方式也十分经济，适合拥有小型网络的企业用户。

2)地理位置分散的网络

虽然网络规模较大，但在地理分布上分散，通过存储域网络存储(Storage-Area Network，SAN)或网络连接存储(Network-Attached Storage，NAS)在它们之间进行互联非常困难，此时可以将各分支机构的服务器采用直接连接存储方式，这样可以进一步降低成本。

3)特殊应用服务器

在一些特殊应用的服务器上，如集群服务器或某些数据库使用的原始分区，均要求存储设备直接连接到应用服务器，可以采用直接连接存储方式。

网络连接存储

- NAS
- 独立于服务器，单独为网络数据存储开发了一种文件服务器，自己形成了一个网络。
- 无需网络文件服务器，不依赖通用操作系统，采用专门用于数据存储的简化操作系统，提高了文档服务效率，响应速度快，数据传输速率高。
- NAS使用TCP/IP协议通信，以文档方式进行数据传输，采用标准文件共享协议：
  - 网络文件系统NFS
  - 超文本传输协议HTTP
  - 公用因特网文件系统CIFS

商用NAS系统

¥2700.00 包邮 45人付款

synology群晖DS218+私有云nas存储网络存储器个人家用共享硬盘盒2盘位私人云盘

华夏存储专营店 北京

掌柜热卖 广告

¥899.00 包邮 53人付款

nas云存储华芸AS1002T V2私有云nas网络存储nas服务器个人2盘位nas主机家用nas

asustor旗舰店 上海

掌柜热卖 广告

¥2700.00 包邮 37人付款

顺丰+五仓发货Synology群晖nas存储DS218+ 家用网络存储NAS企业级主机服务

鼎益数码专营店 河南 郑州

掌柜热卖 广告

¥2700.00 包邮 157人付款

Synology群晖DS218+nas存储服务器主机网络数据家用个人私有云盘存储企业级办公

synology群晖鼎壹专卖店 山东 烟台

¥2700.00 包邮 148人付款

Synology群晖DS218+企业级服务器NAS网络云存储网盘家用私有云盘

synology群晖谷得专卖店 上海

¥2699.00 包邮 6人付款

QNAP威联通TS-532X四核心私有云双万兆网络储存服务器NAS

qnap存储旗舰店 上海

¥4500.00 包邮 2人付款

Synology群晖DS918+网络存储器NAS主机存储服务器Synology私有云存储群辉企业

苏宁易购官方旗舰店 江苏 南京

¥1899.00 包邮 6人付款

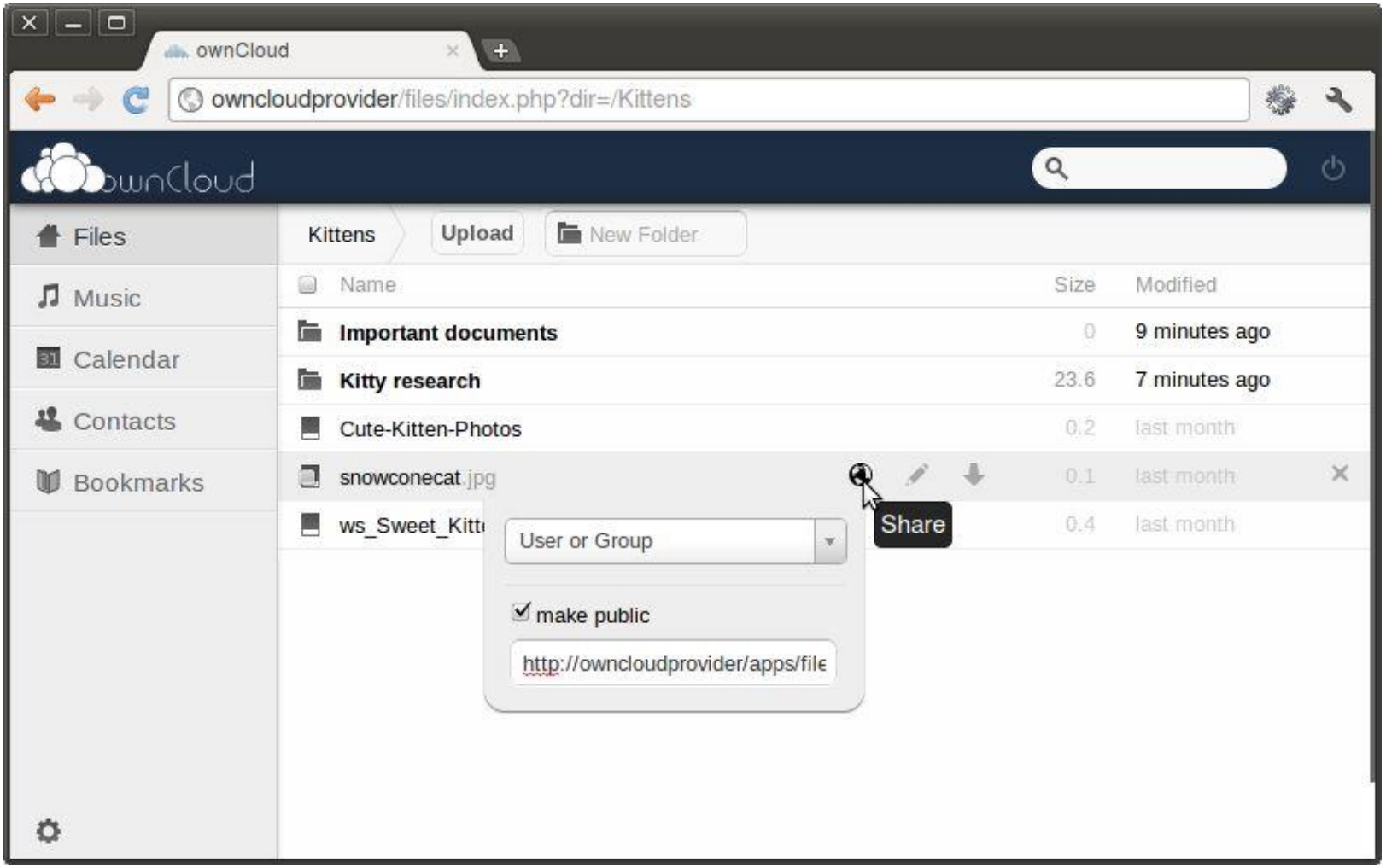
【3期免息】QNAP威联通TS-551双核心4K转码家用商用五盘位磁盘阵列网络存储服务

苏宁易购官方旗舰店 江苏 南京

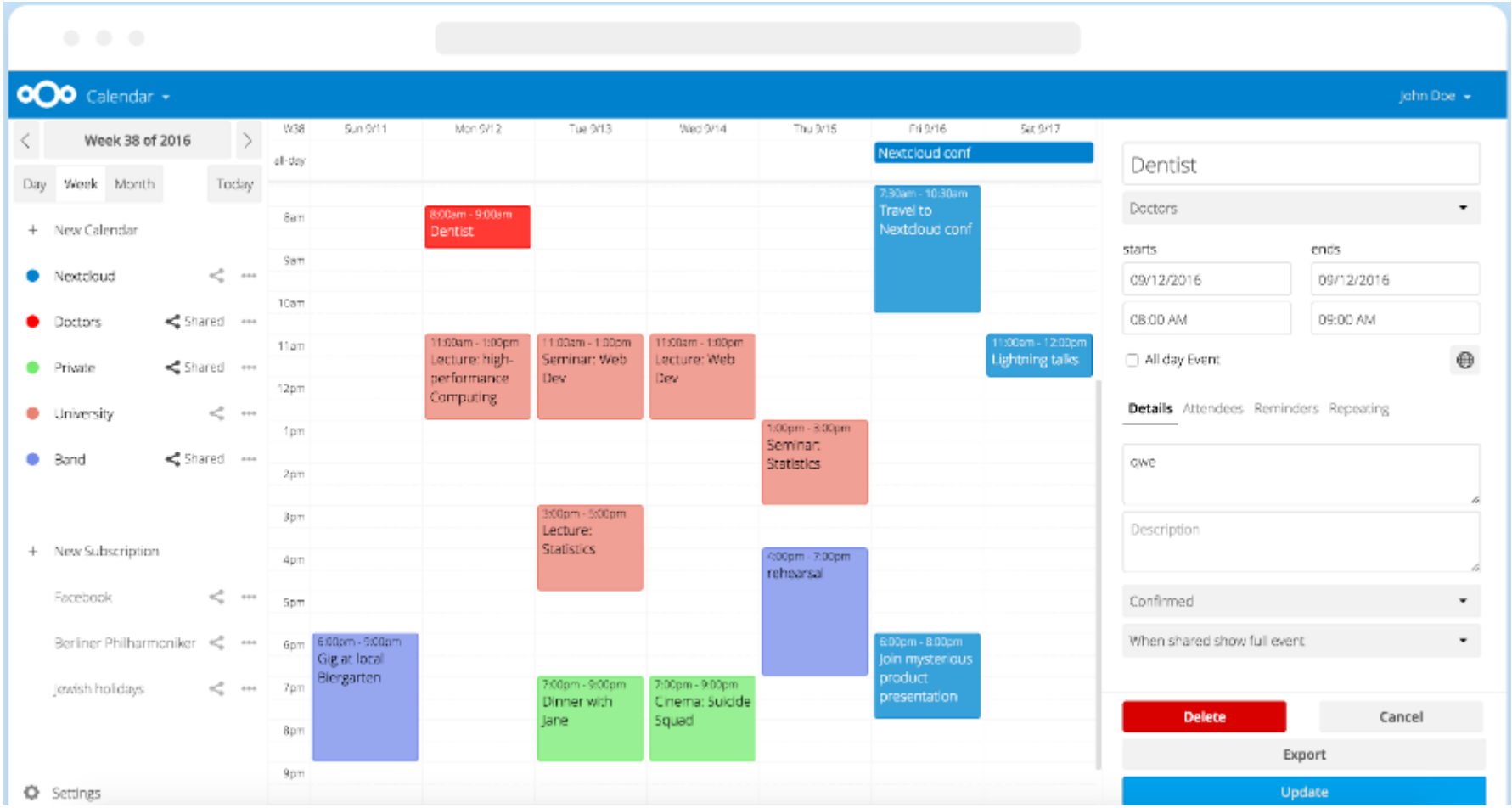
开源网盘

- owncloud
  - https://github.com/owncloud
  - 类似于Dropbox的提供跨平台访问功能的云盘
- Nextcloud
  - https://github.com/nextcloud/
  - 老牌云盘产品
- ibarn
  - https://github.com/zhimengzhe/iBarn
  - 支持fastDfs分布式文件系统
  - KODExplorer
  - 在线编辑
- seafile
  - https://github.com/haiwen/seafile
  - 集群部署性能好，基于ceph网络

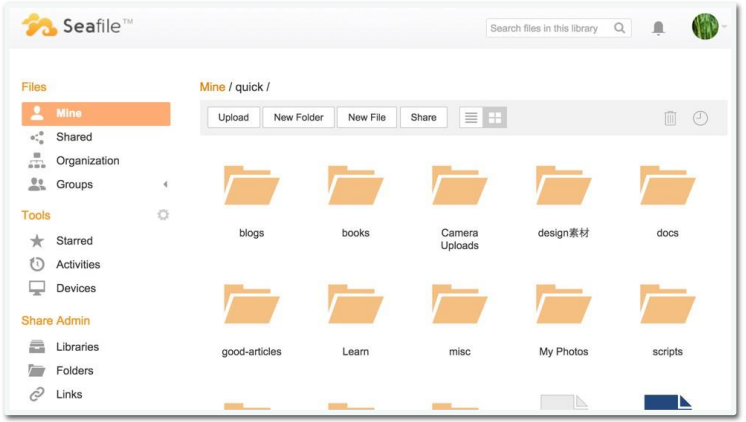
owncloud


ownCloud
owncloudprovider/files/index.php?dir=/Kittens
ownCloud
Files
Music
Calendar
Contacts
Bookmarks
Kittens
Upload
New Folder
Name
Size
Modified
Important documents
0
9 minutes ago
Kitty research
23.6
7 minutes ago
Cute-Kitten-Photos
0.2
last month
snowconecat.jpg
0.1
last month
ws_Sweet_Kitt
0.4
last month
Share
User or Group
make public
http://owncloudprovider/apps/file

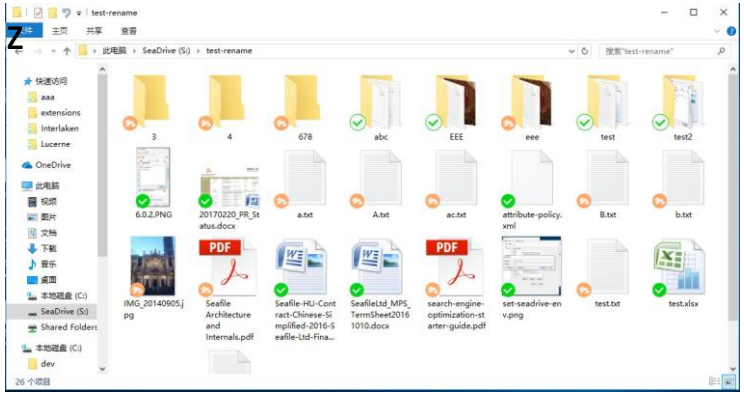
Nextcloud


Calendar
john Doe
Week 38 of 2016
Day
Week
Month
Today
New Calendar
Nextcloud
Doctors
Shared
Private
Shared
University
Band
Shared
New Subscription
Facebook
Berliner Philharmoniker
Jewish holidays
Settings
W38
Sun 9/11
Mon 9/12
Tue 9/13
Wed 9/14
Thu 9/15
Fri 9/16
Sat 9/17
all-day
Nextcloud conf
8am
8:00am - 9:00am
Dentist
9am
10am
11am
11:00am - 1:00pm
Lecture: high-performance Computing
11:00am - 1:00pm
Seminar: Web Dev
11:00am - 1:00pm
Lecture: Web Dev
11:00am - 12:00pm
Lightning talks
12pm
1pm
1:00pm - 3:00pm
Seminar: Statistics
2pm
3pm
3:00pm - 5:00pm
Lecture: Statistics
4pm
4:00pm - 7:00pm
rehearsal
5pm
6pm
6:00pm - 9:00pm
Gig at local Biergarten
6:00pm - 8:00pm
Join mysterious product presentation
7pm
7:00pm - 9:00pm
Dinner with Jane
7:00pm - 9:00pm
Cinema: Suicide Squad
8pm
9pm
7:30am - 10:30am
Travel to Nextcloud conf
Dentist
Doctors
starts
ends
09/12/2016
09/12/2016
08:00 AM
09:00 AM
All day Event
Details
Attendees
Reminders
Repeating
owe
Description
Confirmed
When shared show full event
Delete
Cancel
Export
Update

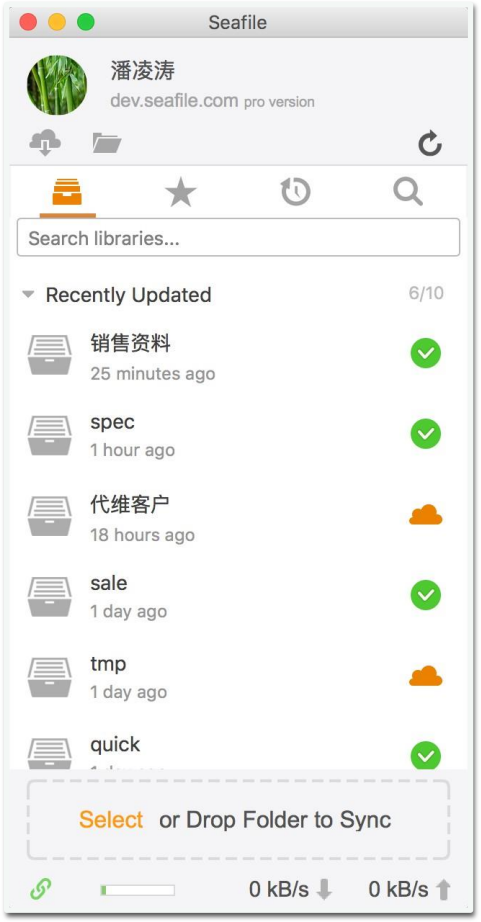
seafile客户端



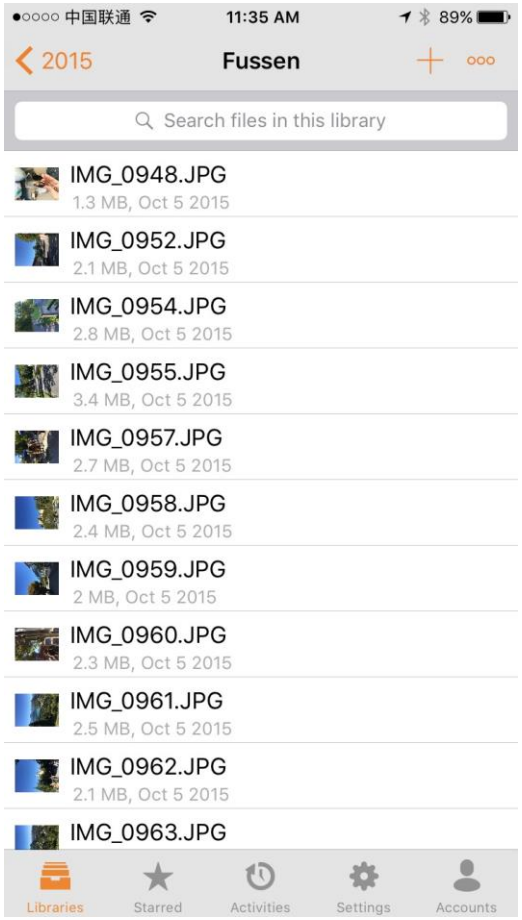
网页客户端



挂载盘

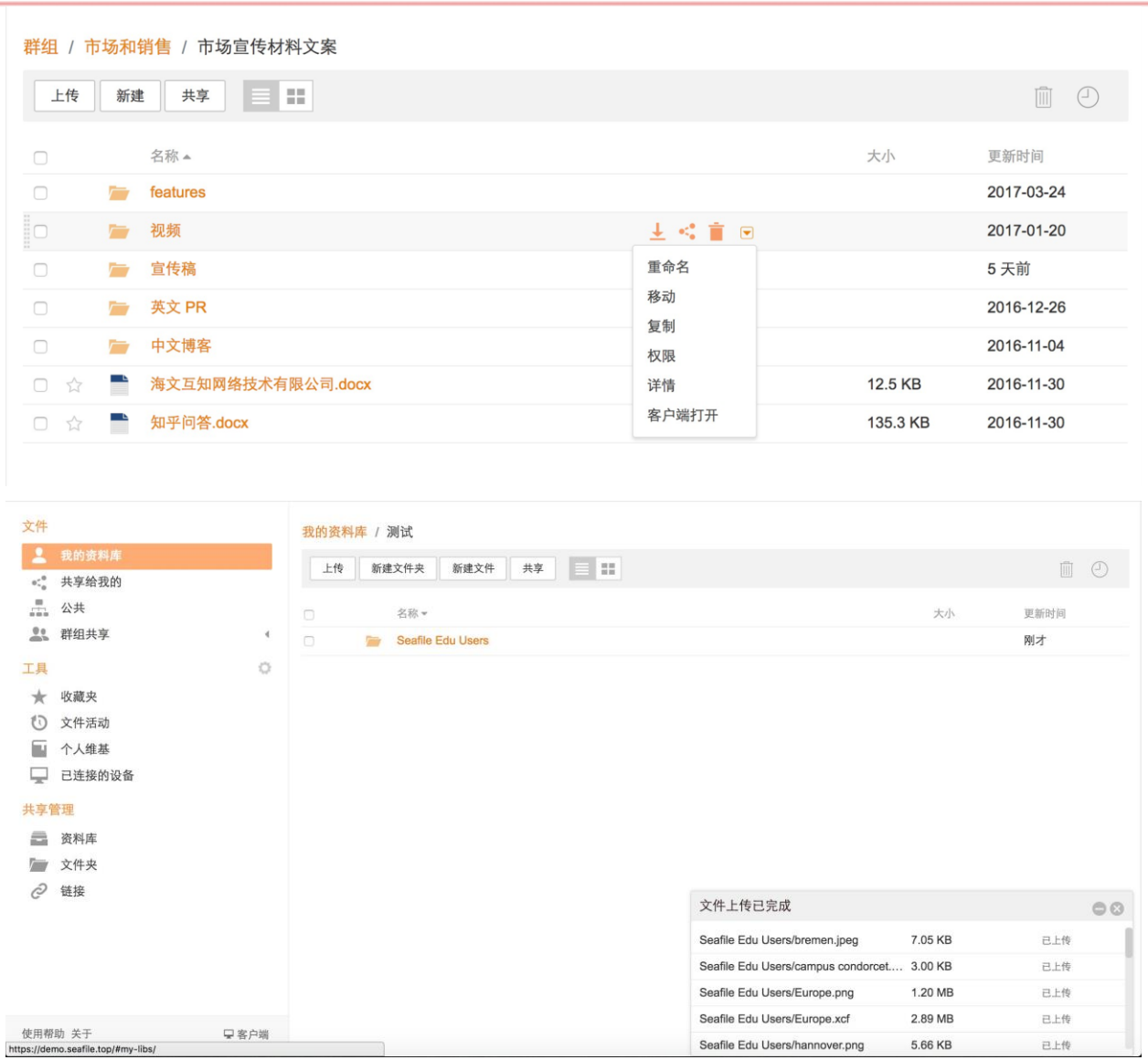


桌面同步客户端

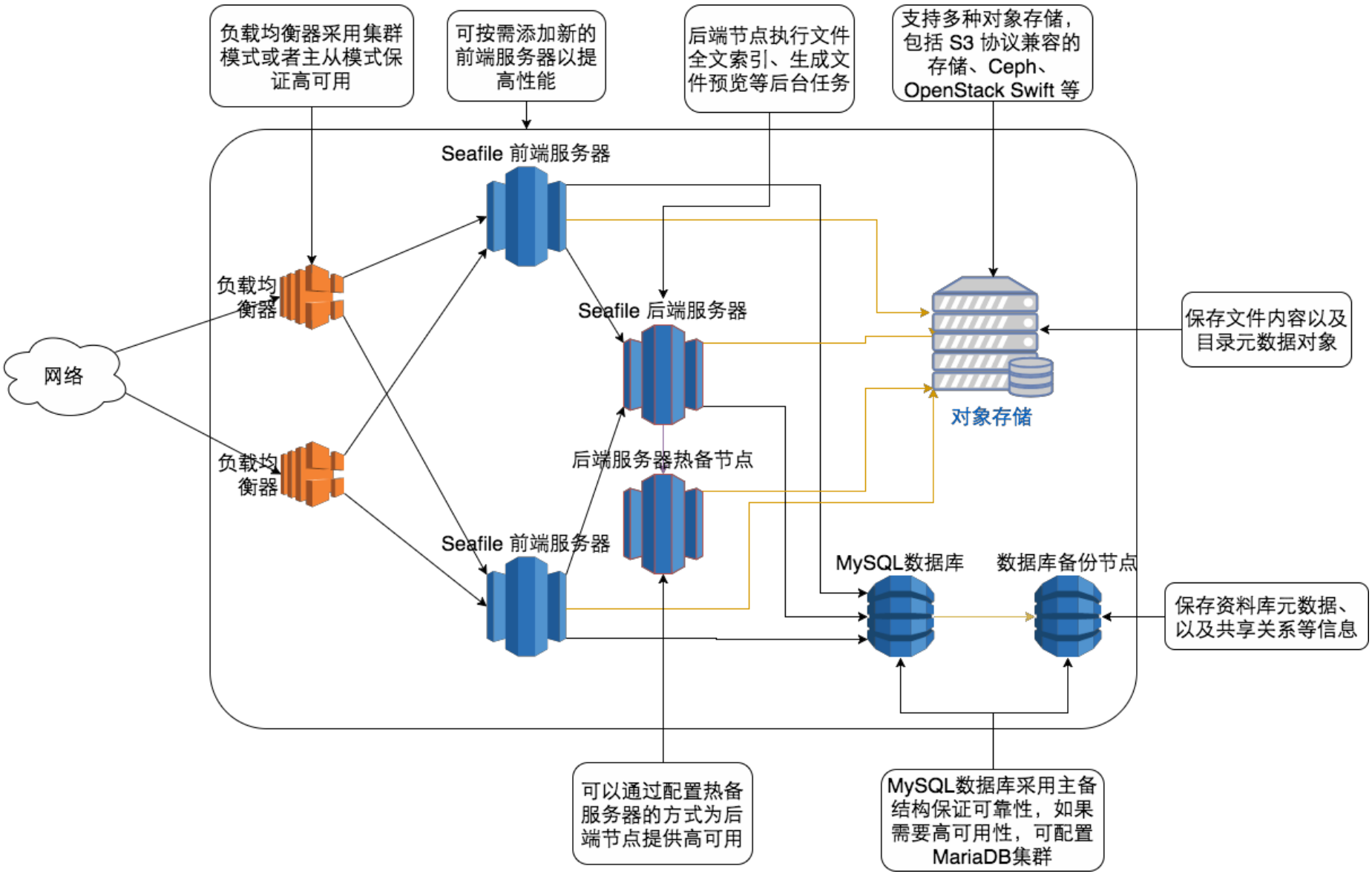


移动客户端

文件上传下载


群组 / 市场和销售 / 市场宣传材料文案
上传
新建
共享
名称
大小
更新时间
features
2017-03-24
视频
2017-01-20
宣传稿
5 天前
英文 PR
2016-12-26
中文博客
2016-11-04
海文互知网络技术有限公司.docx
12.5 KB
2016-11-30
知乎问答.docx
135.3 KB
2016-11-30
重命名
移动
复制
权限
详情
客户端打开
文件
我的资料库
共享给我的
公共
群组共享
工具
收藏夹
文件活动
个人维基
已连接的设备
共享管理
资料库
文件夹
链接
我的资料库 / 测试
上传
新建文件夹
新建文件
共享
名称
大小
更新时间
Seafile Edu Users
刚才
文件上传已完成
Seafile Edu Users/bremen.jpeg 7.05 KB 已上传
Seafile Edu Users/campus condorcet.... 3.00 KB 已上传
Seafile Edu Users/Europe.png 1.20 MB 已上传
Seafile Edu Users/Europe.xcf 2.89 MB 已上传
Seafile Edu Users/hannover.png 5.66 KB 已上传
使用帮助 关于
客户端
https://demo.seafile.top/#my-libs/

seafile部署


负载均衡器采用集群模式或者主从模式保证高可用
可按需添加新的前端服务器以提高性能
后端节点执行文件全文索引、生成文件预览等后台任务
支持多种对象存储，包括 S3 协议兼容的存储、Ceph、OpenStack Swift 等
Seafile 前端服务器
负载均衡器
网络
负载均衡器
Seafile 后端服务器
后端服务器热备节点
Seafile 前端服务器
对象存储
保存文件内容以及目录元数据对象
MySQL数据库
数据库备份节点
保存资料库元数据、以及共享关系等信息
可以通过配置热备服务器的方式为后端节点提供高可用
MySQL数据库采用主备结构保证可靠性，如果需要高可用性，可配置MariaDB集群

存储域网络存储

- 为了克服网络数据传输的瓶颈，引入了SAN储存。
- 通过SAN协议的光纤信道交换机将主机和存储系统联系起来组成一个LUN Based网络，支持多种高级协议。
- 与传统技术相比，SAN能将存储设备从传统的以太网隔离出来，成为独立的存储局域网络。因为采用了网络结构，具有无限的扩展能力。
- SAN采用光纤，数据传输速度高，10Gbps，传输距离可以达到100公里。
- 缺点是成本高，管理难度大。

IP-SAN

- 因为SAN过于昂贵，产生了ISCSI（IP-SAN）
- ISCSI基于IP/TCP协议，通过传统以太网和因特网进行数据传输
- 随着10Gbit以太网络普及，传输速率会进一步提高

7.3 大数据的存储

- 大图数据
- 分布式存储
- 数据管理

大图数据

- 互联网上很多数据都是图数据：
  - 微博粉丝网络
  - 头条新闻推送
- 大图数据的特点
  - 数据规模大
    - PB级别
  - 需要用到分布式内存计算
    - 因为大图数据的基本操作就是图查询

图划分技术

- 负载均衡

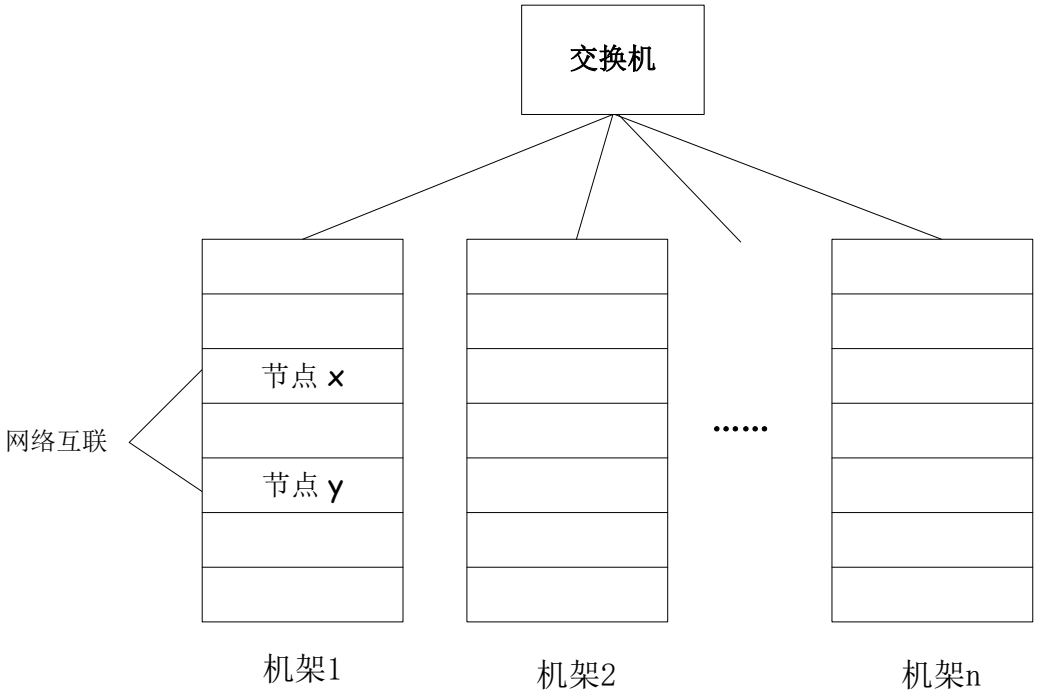
避免网络通信代价的极端方法是将完整的图信息仅存储于一台机器上。显然，这一方式很可能超出单台机器的存储上限，同时这一方法也没有并行计算能力。期望图划分的各个部分具有相近的规模，从而避免负载失衡的情况。负载均衡是相对于机器存储容量和计算能力而言，在复杂的实际应用中，可以构建复杂的度量模型以刻画兼顾机器存储容量和计算能力的负载均衡模型。

- 存储冗余

避免网络通信代价升高的另一极端方法是将图的信息在每台机器上复制一份。但这种方法也容易超出单台机器的存储能力，同时导致大量冗余。对于 k 台机器的分布式系统，这种方法导致 $k-1$ 份存储冗余。为了降低冗余，可以选择特定顶点及其邻接信息进行复制。通常选择度数较大的顶点进行复制，从而在降低通信代价的同时，避免较大冗余。此外，复制顶点个数和位置的选择等都对最终结果有着直接影响。另外，多个副本之间的一致性也是重要的问题，通常需要额外的计算代价保持多份副本的一致性。

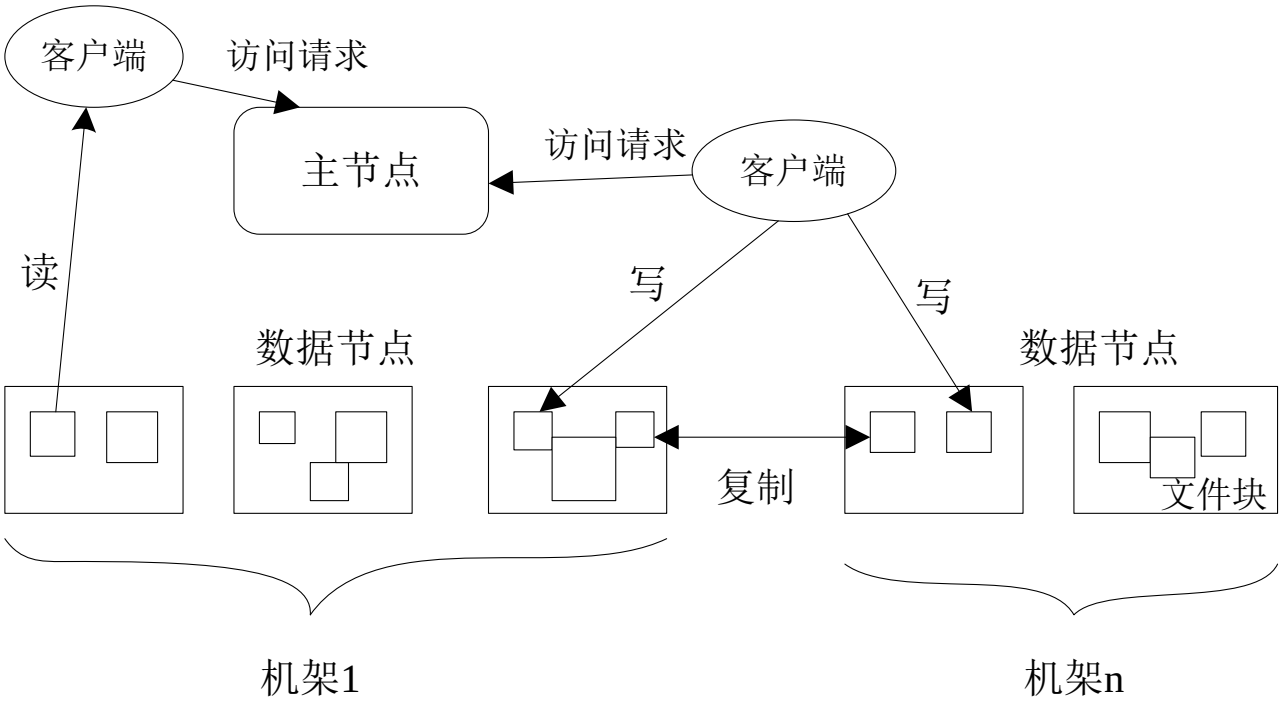
7.4 HDFS

- 分布式文件系统把文件分布存储到多个计算机节点上，成千上万的计算机节点构成计算机集群
- 与之前使用多个处理器和专用高级硬件的并行化处理装置不同的是，目前的分布式文件系统所采用的计算机集群，都是由普通硬件构成的，这就大大降低了硬件上的开销



分布式文件系统的结构

- 分布式文件系统在物理结构上是由计算机集群中的多个节点构成的，这些节点分为两类，一类叫“主节点”(Master Node)或者也被称为“名称结点”(NameNode)，另一类叫“从节点”（Slave Node）或者也被称为“数据节点”(DataNode)



HDFS目标

- 兼容廉价的硬件设备
- 流数据读写
- 大数据集
- 简单的文件模型
- 强大的跨平台兼容性

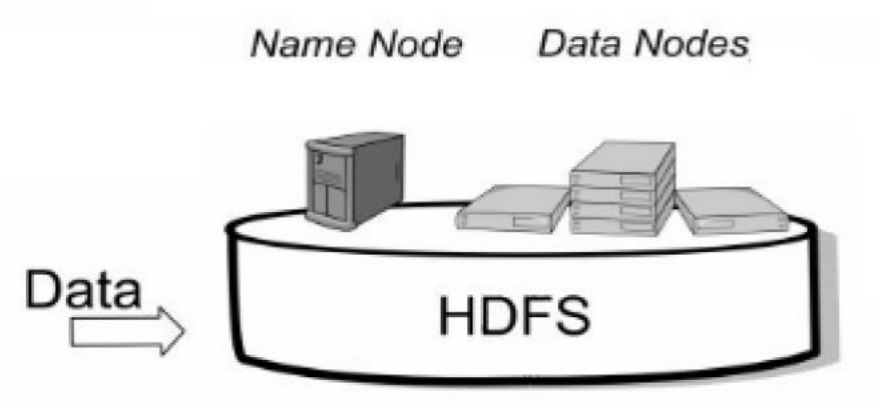
HDFS块

- HDFS默认一个块64MB，一个文件被分成多个块，以块作为存储单位，块的大小远远大于普通文件系统，可以最小化寻址开销，HDFS采用抽象的块概念可以带来以下几个明显的好处：
  - 支持大规模文件存储：文件以块为单位进行存储，一个大规模文件可以被分拆成若干个文件块，不同的文件块可以被分发到不同的节点上，因此，一个文件的大小不会受到单个节点的存储容量的限制，可以远远大于网络中任意节点的存储容量
  - 简化系统设计：首先，大大简化了存储管理，因为文件块大小是固定的，这样就可以很容易计算出一个节点可以存储多少文件块；其次，方便了元数据的管理，元数据不需要和文件块一起存储，可以由其他系统负责管理元数据
  - 适合数据备份：每个文件块都可以冗余存储到多个节点上，大大提高了系统的容错性和可用性

现在还是64MB吗?

- 从2.7.3版本开始，block size由64 MB变成了128 MB的。
- 1.2.1
  - http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Data+Blocks
- 2.5.2
  - http://hadoop.apache.org/docs/r2.5.2/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#Data_Blocks
- 2.7.2
  - http://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#Data_Blocks
- 2.7.3
  - http://hadoop.apache.org/docs/r2.7.3/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#Data_Blocks

HDFS的结构



metadata

File.txt=
Blk A:
DN1, DN5, DN6

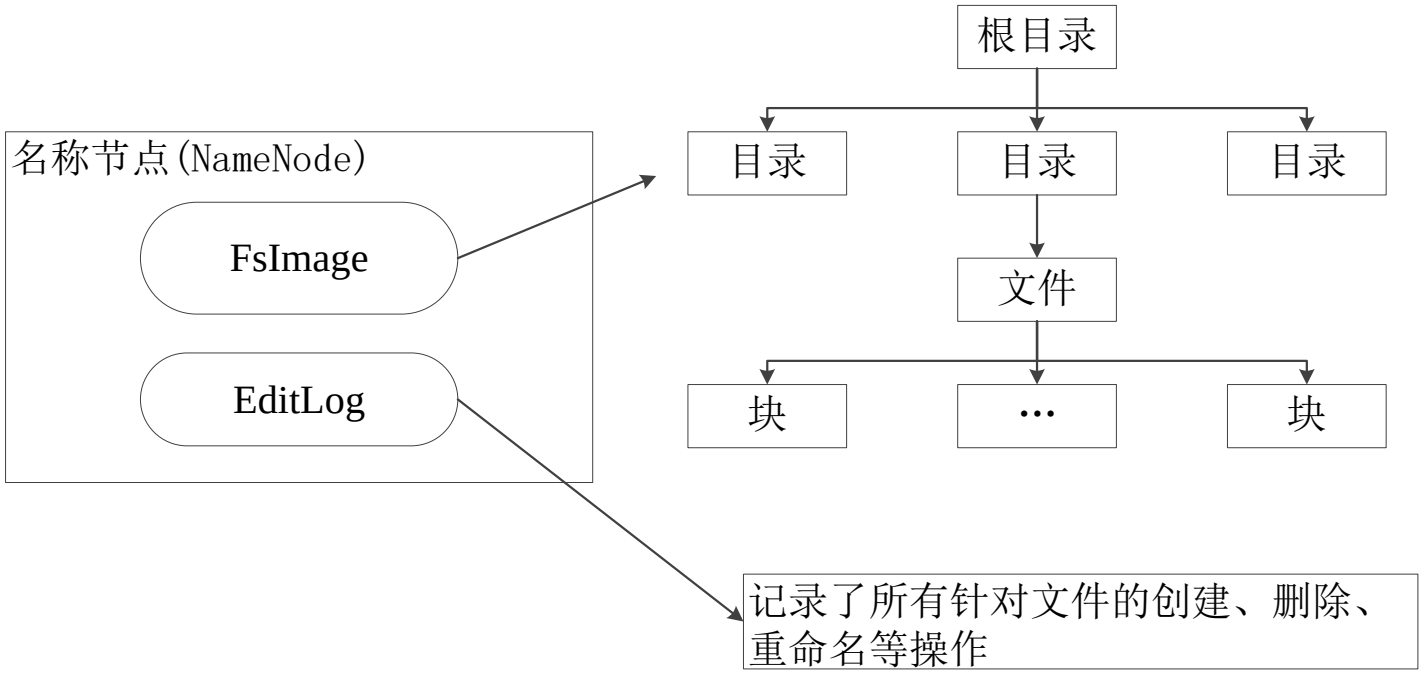
Blk B:
DN7, DN1, DN2

Blk C:
DN5, DN8,DN9

NameNode	DataNode
• 存储元数据	• 存储文件内容
•元数据保存在内存中	•文件内容保存在磁盘
• 保存文件,block， datanode之间的映射关系	•维护了block id到datanode本地文件的映射关系

Namenode的数据结构

- 在HDFS中，名称节点（NameNode）负责管理分布式文件系统的命名空间（Namespace），保存了两个核心的数据结构，即FsImage和EditLog
  - FsImage用于维护文件系统树以及文件树中所有的文件和文件夹的元数据
  - 操作日志文件EditLog中记录了所有针对文件的创建、删除、重命名等操作
- 名称节点记录了每个文件中各个块所在的数据节点的位置信息



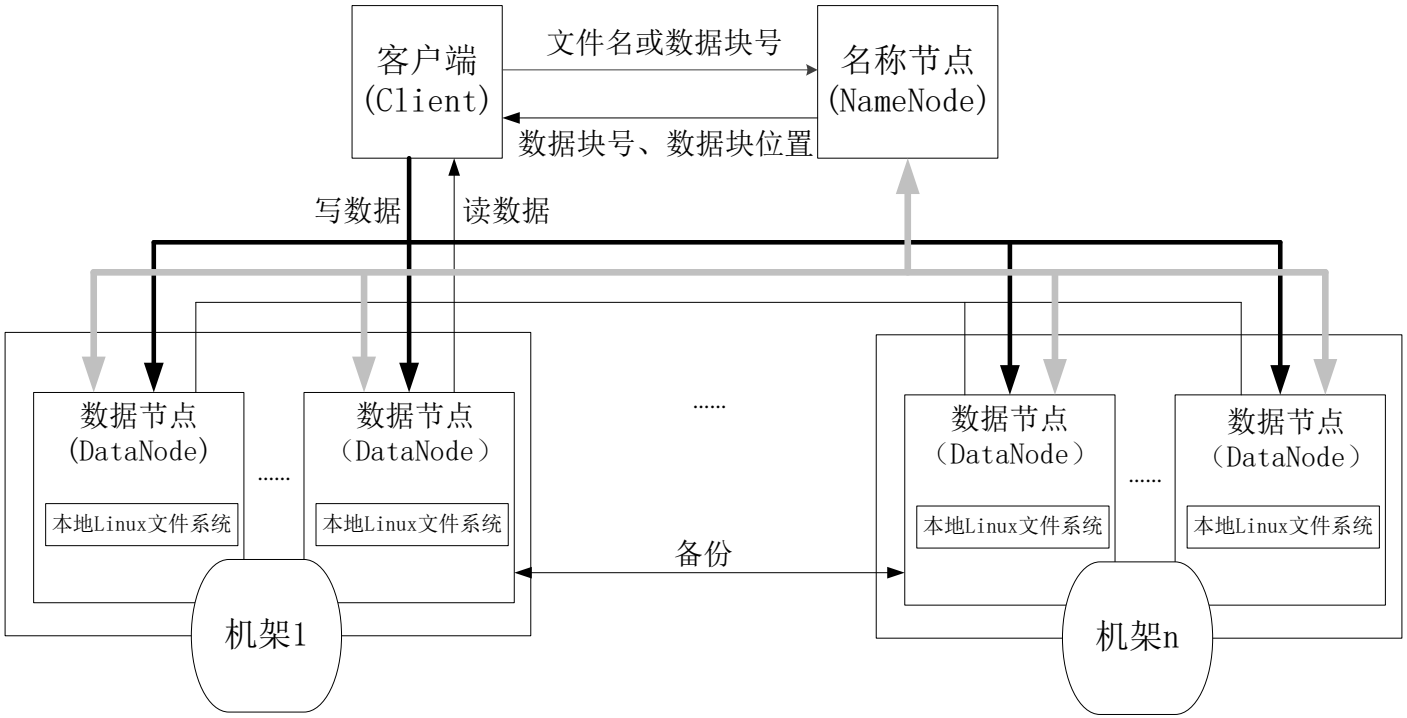
FsImage文件

- FsImage文件包含文件系统中所有目录和文件inode的序列化形式。每个inode是一个文件或目录的元数据的内部表示，并包含此类信息：文件的复制等级、修改和访问时间、访问权限、块大小以及组成文件的块。对于目录，则存储修改时间、权限和配额元数据

- FsImage文件没有记录文件包含哪些块以及每个块存储在哪个数据节点。而是由名称节点把这些映射信息保留在内存中，当数据节点加入HDFS集群时，数据节点会把自己所包含的块列表告知给名称节点，此后会定期执行这种告知操作，以确保名称节点的块映射是最新的。

HDFS体系结构

- HDFS采用了主从（Master/Slave）结构模型，一个HDFS集群包括一个名称节点（NameNode）和若干个数据节点（DataNode）。名称节点作为中心服务器，负责管理文件系统的命名空间及客户端对文件的访问。集群中的数据节点一般是一个节点运行一个数据节点进程，负责处理文件系统客户端的读/写请求，在名称节点的统一调度下进行数据块的创建、删除和复制等操作。每个数据节点的数据实际上是保存在本地Linux文件系统中的



HDFS通信协议

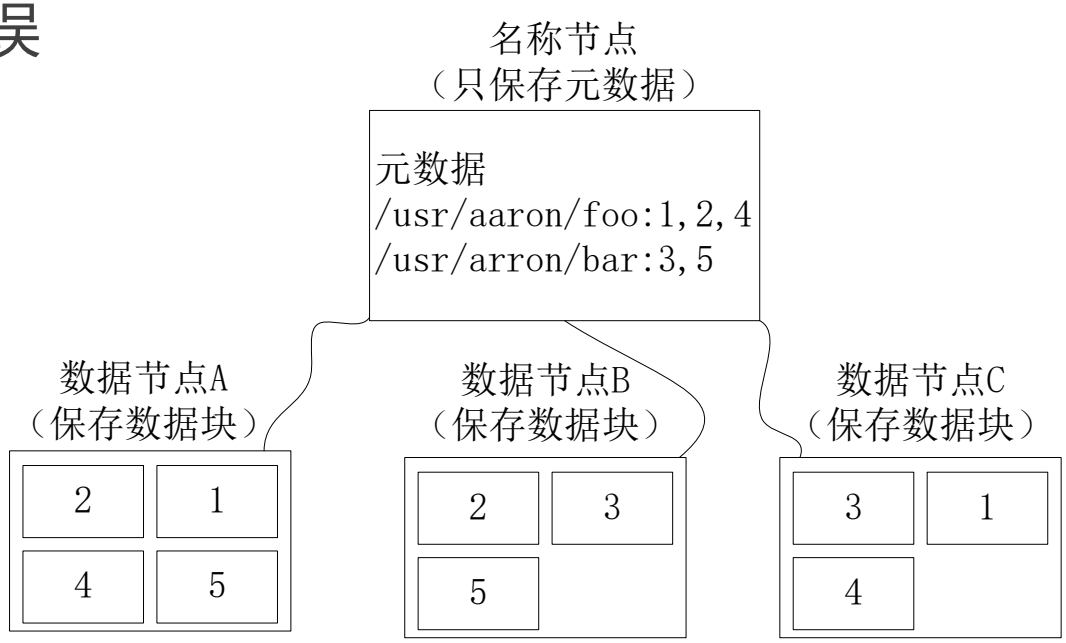
- HDFS是一个部署在集群上的分布式文件系统，因此，很多数据需要通过网络进行传输
- 所有的HDFS通信协议都是构建在TCP/IP协议基础之上的
- 客户端通过一个可配置的端口向名称节点主动发起TCP连接，并使用客户端协议与名称节点进行交互
- 名称节点和数据节点之间则使用数据节点协议进行交互
- 客户端与数据节点的交互是通过RPC（Remote Procedure Call）来实现的。在设计上，名称节点不会主动发起RPC，而是响应来自客户端和数据节点的RPC请求

HDFS客户端

- 客户端是用户操作HDFS最常用的方式，HDFS在部署时都提供了客户端
- HDFS客户端是一个库，暴露了HDFS文件系统接口，这些接口隐藏了HDFS实现中的大部分复杂性
- 严格来说，客户端并不算是HDFS的一部分
- 客户端可以支持打开、读取、写入等常见的操作，并且提供了类似Shell的命令行方式来访问HDFS中的数据
- 此外，HDFS也提供了Java API，作为应用程序访问文件系统的客户端编程接口

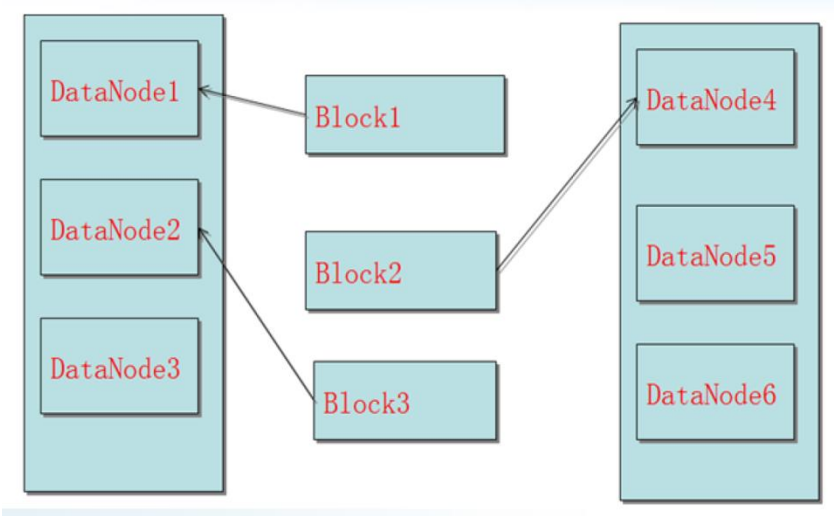
HDFS冗余数据保存

- 作为一个分布式文件系统，为了保证系统的容错性和可用性，HDFS采用了多副本方式对数据进行冗余存储，通常一个数据块的多个副本会被分布到不同的数据节点上，如图所示，数据块1被分别存放到数据节点A和C上，数据块2被存放在数据节点A和B上。这种多副本方式具有以下几个优点：
  - 加快数据传输速度
  - 容易检查数据错误
  - 保证数据可靠性



HDFS数据存取策略

- 1.数据存放
  - 第一个副本：放置在上传文件的数据节点；如果是集群外提交，则随机挑选一台磁盘不太满、CPU不太忙的节点
  - 第二个副本：放置在与第一个副本不同的机架的节点上
  - 第三个副本：与第一个副本相同机架的其他节点上
  - 更多副本：随机节点



HDFS数据存取策略

- 2. 数据读取
  - HDFS提供了一个API可以确定一个数据节点所属的机架ID，客户端也可以调用API获取自己所属的机架ID
  - 当客户端读取数据时，从名称节点获得数据块不同副本的存放位置列表，列表中包含了副本所在的数据节点，可以调用API来确定客户端和这些数据节点所属的机架ID，当发现某个数据块副本对应的机架ID和客户端对应的机架ID相同时，就优先选择该副本读取数据，如果没有发现，就随机选择一个副本读取数据

HDFS数据错误与恢复

- HDFS具有较高的容错性，可以兼容廉价的硬件，它把硬件出错看作一种常态，而不是异常，并设计了相应的机制检测数据错误和进行自动恢复，主要包括以下几种情形：名称节点出错、数据节点出错和数据出错。
- 1. 名称节点出错
  - 名称节点保存了所有的元数据信息，其中，最核心的两大数据结构是FsImage和Editlog，如果这两个文件发生损坏，那么整个HDFS实例将失效。因此，HDFS设置了备份机制，把这些核心文件同步复制到备份服务器SecondaryNameNode上。当名称节点出错时，就可以根据备份服务器SecondaryNameNode中的FsImage和Editlog数据进行恢复。

HDFS数据错误与恢复

- 2. 数据节点出错
  - 每个数据节点会定期向名称节点发送“心跳”信息，向名称节点报告自己的状态
  - 当数据节点发生故障，或者网络发生断网时，名称节点就无法收到来自一些数据节点的心跳信息，这时，这些数据节点就会被标记为“宕机”，节点上面的所有数据都会被标记为“不可读”，名称节点不会再给它们发送任何I/O请求
  - 这时，有可能出现一种情形，即由于一些数据节点的不可用，会导致一些数据块的副本数量小于冗余因子
  - 名称节点会定期检查这种情况，一旦发现某个数据块的副本数量小于冗余因子，就会启动数据冗余复制，为它生成新的副本
  - HDFS和其它分布式文件系统的最大区别就是可以调整冗余数据的位置

HDFS数据错误与恢复

- 3. 数据出错
  - 网络传输和磁盘错误等因素，都会造成数据错误
  - 客户端在读取到数据后，会采用md5和sha1对数据块进行校验，以确定读取到正确的数据
  - 在文件被创建时，客户端就会对每一个文件块进行信息摘录，并把这些信息写入到同一个路径的隐藏文件里面
  - 当客户端读取文件的时候，会先读取该信息文件，然后，利用该信息文件对每个读取的数据块进行校验，如果校验出错，客户端就会请求到另外一个数据节点读取该文件块，并且向名称节点报告这个文件块有错误，名称节点会定期检查并且重新复制这个块

HDFS命令行操作

>hadoop fs –help 查看帮助

上传文件

>hadoop fs –put localfile hdfs_path

>hadoop fs –copyFromLocal localfile hdfs_path

创建文件夹

>hadoop fs –mkdir /data

删除文件或文件夹

>hadoop fs –rm –r –f /data/file

创建一个空文件

>hadoop fs –touch /data/nullfule

在web上访问hdfs

- 关闭防火墙：service iptables stop
- 打开浏览器输入：
- http://192.168.17.200:50070/explorer.html#/

Browse Directory

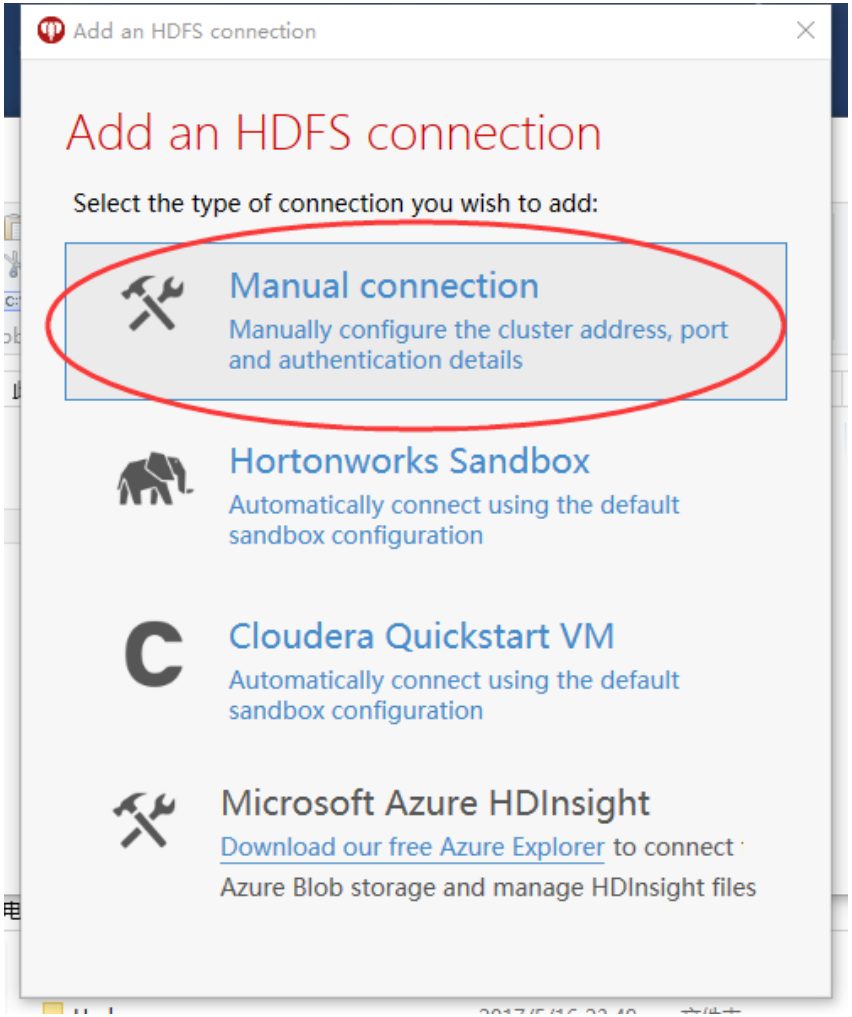
/ Go!

Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	zzti	supergroup	4 B	1	128 MB	aaa
drwxr-xr-x	zzti	supergroup	0 B	0	0 B	in
drwxr-xr-x	zzti	supergroup	0 B	0	0 B	out
drwx------	zzti	supergroup	0 B	0	0 B	tmp
drwxr-xr-x	zzti	supergroup	0 B	0	0 B	zhangsan
drwxr-xr-x	zzti	supergroup	0 B	0	0 B	zhangsan_out

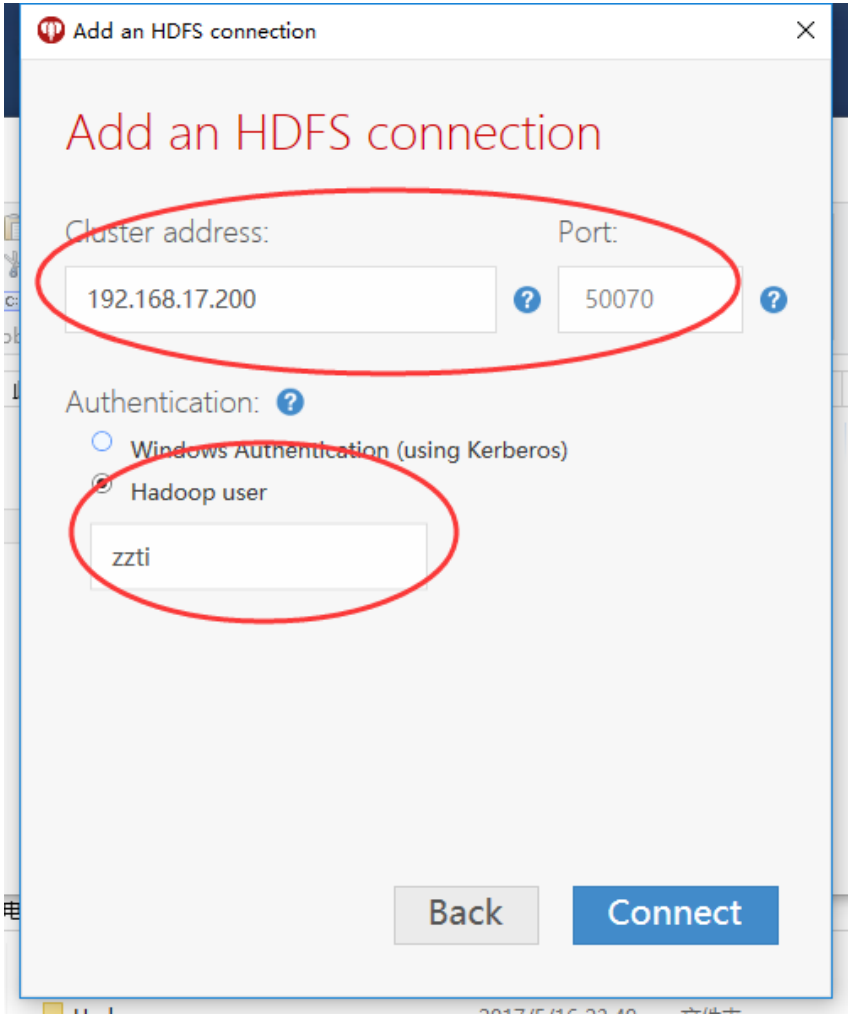
如何在windows上访问hdfs?

- HDFS Explorer是一个在windows上管理HDFS系统的工具，支持上传、下载、重命、复制、移动和删除等。

HDFS Explorer



HDFS Explorer



HDFS Explorer