# 分布式存储的设计与实现

摘要：目前随着互联网的发展，用户，企业已经实现了资源的共享，共享的内容包括了计算能力，存储能力，网络传输，数据服务等。随即出现了云服务，相比较而言，云服务比个体服务的搭建有诸多的优势，如果一个中小企业需要相应的云服务，通过自己构建机房搭建平台，无疑不是个明智的选择，因为一套可以上线的服务需要付出巨大的开发成本和维护成本以及硬件成本，明显已经超过了服务所能创造的价值，当互联网共享云服务出现了，给企业带来了很大的便利，用最少的成本创造最大的价值。

云平台主要使用Hadoop开源组件，首先Hadoop实现了一个分布式文件系统，具有较高的容错能力和高可用性能，云存储主要使用了Hadoop的HDFS海量数据存储的能力，将许多廉价的机器经过整合管理变成具有超大处理性能的云存储系统，在互联网中，云存储是广大客户可以共享的一种存储服务，面对的客户大到企业，小到个人，提供的服务具有较好的安全性，可靠性，易管理性等等优点；Hadoop是通过java语言实现，我们在开发过程中可以通过java应用程序调用相关的API，并封装成相应的功能，实现云存储服务。

本研究的主要意义在于如何将现有的分布式存储运用到实际应用当中，在世纪生产环境当中，仍然有很多从云存储基本功能基础上延伸出来很多的新功能，比如ftp服务器对接云存储，这项功能主要是让原有的ftp服务器底层对接windows系统，通过该写相应的特性，让底层的文件系统对接云存储文件系统，而对于客户来说，仍然觉得云存储文件系统和windows文件系统一样，由文件和子文件以及属性一样，实际上在云存储中也不存在云存储文件系统，这是就需要约定一套规定来对云数据进行翻译，从而实现和客户进行完美的交互；其次是对于海量数据的存储，如何设计性能更好，更符合生产环境下的存储系统。

关键字：云存储，Hadoop，HDFS，ftp。

／／xie

第一章引言

1.1背景研究

1.1.1背景介绍

大数据时代，我们需要存储的数据已经急剧增长，企业以及互联网的数据都在以每年50%的速率在增长，经过预测，到2020年，全世界的数据已将达到35ZB，等于80亿块4TB的硬盘，面对如此庞大的数据，我们的存储结构将面临巨大的挑战，这时非结构化数据存储将占很大一部分比例。互联网的发展，让数据产生的主体由企业转向了个人，对于个人来说，产生的很多数据都是图片，文档，视频等非结构化数据，这些数据都需要使用非结构化数据库来存储。对于这种分布式存储，物理介质分不到不同的地理位置，数据上传，可以进行就近上传，数据存盘时，根据映射算法，找到合适的IP，把数据刷到指定的磁盘。这种分布式数据存储，有效地解决了因为数据存储和访问造成的IO瓶颈，下载数据的时候，可以同时从多个数据盘上去取数据，上传数据的时候能够高并发存储，从单个数据流存储数据上看，映射算法会使用到负载均衡原理，有效地选择经验值最高的IP进行存储，大大提高了存储的效率。而且随着数据量的增加，我们需要从硬件上去扩大存储能力，这在集中式存储系统中是很难做到的，换更大的磁盘和内存盘，从硬件上来说，过程繁琐，操作艰难，从软件层面上来说，数据的转移相当困难，经济上也会付出更大的代价，大容量单盘比小容量多盘的成本高，于是分布式存储给使用者带来了福音，当我们需要增加平台的存储能力时，我们仅仅需要增加一台设备放到局域网中，通过管理软件Zookeeper的API把新的设备添加到集群中，并启动相应的负载均衡功能，让新的设备均摊集群承载的数据量，这些优良的性能和数据时代的大环境让分布式存储的盛行成为了必然。

1.1.2相关课题介绍

目前云存储平台主要是给其他应用程序提供数据的最底层的数据存储和数据产查询的服务，其实就是数据库的分布式实现，对外提供相应的接口供开发人员访问，对于本平台的研究的同时，我们还解决了对接云存储数据的ftp服务器，底层采用的是FTP协议，顾名思义就是文件传输协议，实现服务端的文件共享，在互联网上的计算机可能会使用不同的操作系统，计算机Unix，Dos，Windows，MACOS等等，为了实现不同系统之间的文件交流，我们就需要建立一个统一的文件传输协议，这就是FTP协议，这样用户就可以从别的用户那里获得文件，或者把文件传送给其他人。目前FTP主要采用了客户机／服务器系统，在服务端搭建一个提供文件共享的服务进程，用户通过一个支持FTP协议的客户端程序，连接服务端，发出相应的指令，和服务端实现具体的交互。本课题研究相关的FTP服务器就是将底层的数据存储对接在云端，对于云端，数据都以流的形式存在，这时候，文件系统的管理需要由HBASE来管理，每个流对应一项元数据，包括这一分片的大小，序列值，存储路径等等相关的属性，而对于真实数据流则直接实现刷盘，存在磁盘的目录树中，这样便实现了数据的高效率存储和管理，同时也充分利用了HBASE的列可以动态增加，如果列空，则不占用空间，同时，列数据增加，会进行自动的切割，使数据可以水平扩展，同时HBASE可以提供高并发读写操作的支持。

1.2研究现状

1.2.1国外研究现状

目前国外的云存储已经有一定的成功案例，首先Amazon公司是最先研究关于云存储系统的企业，目前在存储体系上也是最完备最成功的，云存储是一项能够给广大用户提供服务的平台，对于Amazon公司来说，它拥有远远超过它本身所需要的硬件资源，于是在06年将存储服务对外放开，作为AWS的一部分。目前Amazon提供的云存储服务价格低廉，性能稳定，是业界最受欢迎的存储服务，目前Amazon的简单存储服务支持上传的单个文件不可以大于5T；IBM在09年也推出了企业级智能云存储服务，主要通过存储虚拟化和基于私有云的存储技术，目前在提供应用程序方面得到了广泛的支持，每个企业的数据管理模式存在一定的差别，IBM可以根据企业的本身框架进行调整，通过虚拟化和自动化技术，构建属于企业自身的云存储系统，充分协调企业自身的硬件和软件资源，让资源统一分配，管理，部署，备份和监控，更加有效地实现资源的共享；EMC公司拥有一套基于策略的存储管理系统ATMOS，这是一套首次能把容量提高到PB字节的存储方案，通过全世界的云存储环境，把客户的那个大量的非结构化数据自动管理起来，并且为其提供自动的信息配置功能，使客户安全地构建和实现云存储服务，ATMOS的优势在于可以提供信息配送和处理，使用一套策略来建立云存储的不同层级，目前主要容量为120T，240T和360T；2010年夏天Google提出了GSD服务，该服务实现了数据可以在美国的若干个数据中心之间复制数据，组织方式提出了Bucket的概念，而且Google向外提供了管理工具和网络用户界面，完全放开存储服务；为了应对云存储的发展，HP提出了云存储融合架构，将网络，软件，资源和服务器统一管理起来，形成资源池，并通过模块化存储来实现容量和性能的提高。

1.2.2国内研究现状

目前国内的各大高校都在积极开展云存储的相关方案的研究，首先对于清华大学的Corsair系统，实现了分布式存储和数据共享，主要为学校师生提供个人数据存储，社区资源共享和公共资源的下载等，这个存储平台目前有几百个集团用户和一万六千多个人用户，大概使用100T左的存储空间；解放军理工大学的Mass云，这个云平台陈本较低，性能高校，在存储较多数据量的时候，性价比较低，而且通用API方便开发人员进行二次开发，使用成本较低。

国内的产业界也纷纷推出自己的云存储方案：

网易NOS深度整合了IaaS，PaaS以及各种容器技术，服务还包括了弹性计算Devops工具链和微服务基础等服务；腾讯COS对于开发者来说非常方便，性能安全，稳定，高效，访问主要是基于HTTP的方式进行访问；青云QingStor 对象存储服务提供了无限扩展的存储空间，存储高效，性能可靠；阿里OSS主要适合存放任意文件类型，对企业个性化服务来说是个不错的选择；华为云最大的特点是在于IT基础设施资源能够随着用户使用情况作弹性伸缩，而且华为云是经过行业认证和授权的专业平台，网络接入到配备管理具有7层安全防护，具有极高的稳定性和安全性。

1.3平台主要内容和结构

1.3.1实验需要解决问题

一直以来，我们做项目使用的都是关系型数据库，存储也采用的集中式，对于关系型数据库，处理起来非常方便，而且使用集中式存储一般来说硬件资源几种一个活着几个设备上，但是随着数据量的增加，动态资源的调整，这种集中式存储越来越不能满足我们的需求，实验的数据是实时从设备上采集而来，数据量也越越来越大，读数据的访问也越来越频繁，而且数据的安全性和扩展性要求也越来越高，这时候集中式的关系行数据库已经远远不能满足我们的需求，于是出现了分布式存储，将组件分布在网络计算机上，很多台计算机一起为外部提供服务，而且在同一个分布式系统中，计算机在空间部署上可以随意分布的，这些计算机可以放在不同的机柜，不同的机房甚至是不同的城市，因此这样的分布式系统基本没有逻辑上的约束；由于单节点的安全性得不到有效的保证，而分布式存储实现了三幅本保存，当一个节点的数据丢失的时候，可以通过心跳得知，并通过Namenode组织发出任务，从其他的副本中复制数据完成形成丢失的副本；除此之外，随着访问量的增加，数据访问出现了瓶颈IO，对于分布式系统，对于资源的请求可以实现多线程高并发访问，同时从不同的设备上访问多项数据，大大增加了系统的访问性能，从以上这些特性来看，分布式存储将会给我们的应用服务带来更加稳定高效的支持。

1.3.2论文主要研究内容

1.3.3论文的结构

第2章 云存储的介绍

21定义

云存储是从云计算中发展而来的一个提供网络存储技术，云存储涉及集群系统，网络技术和分布式文件系统，以往的单节点存储已经不能满足时代的需求，随即出现了一套能够将网络中的各种各样不同类型的存储设备管理起来，提供对外的数据存储和业务访问功能。它不是单一的某台设备，而是一整套完整的服务，通过分布式技术把一个个小的存储设备转化成对外提供的服务，这套服务包含了网络设备，存储设别，服务器，应用软件，公共访问接口，接入网和客户端程序等等多个部分。

22优势

有了云存储，传统的集中式存储在很多行业已经渐渐被代替了，云存储有着传统存储不可代替的优势。

221可扩展性

云存储的可扩展性，目前大多数企业在业务刚刚兴起的时候，很难预估未来的业务扩展量，不管是计算需求还是存储需求，增长速率也不可预测，这时候对存储服务的扩展性要求很高，对于云存储来说，集群是通过Zookeeper来管理的，Zookeeper提供对外API，能够方便快捷通过API调用相应的进程，把新的设备IP添加进来，这台设备只需要安装相应的操作系统和应用软件即可，并在网络中运行，这种动态容量扩充非常方便，不需要对原有的服务造成任何的影响，通过内部封装好的均衡负载功能，将新加进的设备均衡负载整体的存储量。

222安全管理

整个存储系统是一大批设备，传统的存储模式对于设备的管理尤其困难，对于每个设备的使用情况，各种指标的指数，尤其是当服务器或者磁盘出现故障了，这时候对数据的读取和写入会受到很大的影响，甚至造成数据的丢失，有了云存储久不存在这样的问题了，所有的数据都存在多个副本，当某个磁盘的数据发生丢失，这是可以通过心跳感知设备的异常，当发现异常，便可以调动相应的程序自动从其他的盘中拷贝数据形成新的副本，这样给管理人员带来了很大的便利，减轻了工作负担，大大提高了存储系统的维护效率；同时在进行设备升级的时候，可以在软件层面上操作，把旧设备上的数据进行拷贝转移，这样就可以在提供稳定服务的同时卸下旧的设备，把新的设备替换上，再将数据迁移回来，实现动态升级且保持服务的连续性。

223 设备类型广泛

目前设备的更新换代尤其快，一种新的设备在出厂后2年可能就被淘汰停产了，很难再买到相同型号的设备，以往的存储系统中，对磁盘的一致性要求非常高，在容量，厂家，型号方面必须完全一样，不然一起运行过程中就会出现不兼容的现象，这在互联网行业是一个极大的弊端，互联网行业是个业务不断复杂化，数据量不断增加的一个行业，这是我们就需要充分考虑增加业务量时付出的代价，对于非集中式存储，首先我们需要把数据拷贝到新的设备上，同时单台设备的性能提升需要付出高昂的代价，相对而言，相同的性能，多台设备比单台设备要廉价很多，而且添加新设备的同时，我们不需要限制新设备的规格，这样就可以选择大量性价比较高的设备，降低服务的成本，新增设备的时候，旧的设备仍然能够提供稳定的服务。

23应用场景

目前可以说云存储的应用场景比比皆是，典型是音视频应用场景，首先对于音视频点播类应用场景，对服务的性能要求比较高，播放需要非常流畅，而且在音视频处理上也有着很大的需求，对于音视频文件在服务端是需要进行相关的转码操作，有些类型的视频适合在web端播放，有些视频适合在移动端上传，这时我们需要把食品在服务端进行转码，对于上传的文件通常有些比较大，这时需要采用断电续传的原理保证上传的速度和稳定性，同时为音视频服务提供性能保障的是CDN节点可以提供缓存功能，对于某一点的高频访问数据，可以放在缓存中，这里的缓存有2层含义，首先对于单点访问，一级频繁访问数据会放在周围边缘节点的内存中，调用的速度最快，其次二级频繁访问数据会放在边缘节点的磁盘上，层层上升，节点数据量也越来越大，数据越来越全；数据传输过程中，数据会暴露在网络中，对于私有资源会进行加密，访问数据时需要下载相应的凭证，保证资源的私密性。

24 存储结构

云存储系统平台从顶层到底层主要包含了4个层次：

数据存储层、数据管理层、数据服务层、和用户访问层。

241数据存储层

首选数据存储层是服务平台的最底层进行数据处理的层，这层直接接触的是存储设备和网络设备，对于这些设备还对应了一个用来管理的系统，将这些硬件设备同意管理起来，包括设备的使用情况，设备的性能等等，以及设备的维护升级。整个云存储系统可以对外提供不同的存储服务，对于不同的数据类型，统一放在一个海量的存储池中。所有的数据都是对象，文件的结构由HBASE来管理，真实数据存在磁盘目录中，所有的数据的查找都是通过IP和对象的目录来查找的，一个文件的所有数据可能存放在不同的设备，不同的磁盘中，目前单服务器肯定是难以满足高并发访问的需求，P2P的数据结构要求比较高，在网络中需要大量的节点和科学的算法来保证数据传输的可靠性和质量。因此这种多元化存储方法能够解决这样的问题，构建不同地区的数据中心，在不同的区域可以访问周边的边缘节点的数据，当用户数量增大时，可以缓解中心服务器的IO压力，为客户带来更好的服务。

242数据管理层

数据管理层是整个云存储的核心，也是最复杂的部分。该层一般都是通过集群管理，存储方式通过分布式存储。在这一层中针对用户具体的需求，作出相应的扩展性功能，对一些私密性较强的数据，可以对数据进行加密，加密有不同的加密方式，针对不同的数据类型，可以采取不同的加密方式，以及备份和冗余等等。存储层层包含了很多重要的功能，对上层提供的服务也是各种各样，如果不进行有效的管理，上层调用数据层的功能很不方便，通过数据管理层的包装，可以把各种功能进行统一的管理，诸如用户管理，安全管理，副本管理以及策略管理等来实现对公共数据的统一管理，如此就屏蔽了存储底层的功能调用方式和实现机制，实现上层和底层的完美对接，让多个存储设备之间能够协调工作起来，更好地提供服务。

243**数据服务层**

这一层是存储服务直接面向客户的一层，在这一层**根**据具体的需求灵活地对服务进行相应的扩展。在这一层，存储平台暴露的接口已经形成了一定的标准，根据用户的不同，可以对接口做不同的实现，最终提供相应的服务。例如果数据存储服务，主要为用户分配相应的用户和空间，提供对象操作，根据用户使用服务的时间和对象操作的次数来进行相关的计费；公共资源服务，提供所有用户对一部分资源的访问可达性；多用户共享服务，比如一个公司的不同部门可能对公司的一部分数据需要共享，这时会根据用户定义共享组；数据备份服务，这里主要是为数据服务提供安全保障，当机器发生故障的时候，系统的可用性能够不受到影响。

244用户访问层

这一层是用户主要是用户能够真实可以接触的一层，关于云存储的相关的功能在这一层已经完全包装好，用户使用的仅仅是客户端或者URL，服务端分配不同用户相应的权限，而用户即可在任意点通过互联网登录系统，通过权限验证，进入存储系统，访问相关已经被授权的服务。

第三章 需求分析

31需求分析

32主要功能分析

33用例图分析

第4章云存储的核心技术

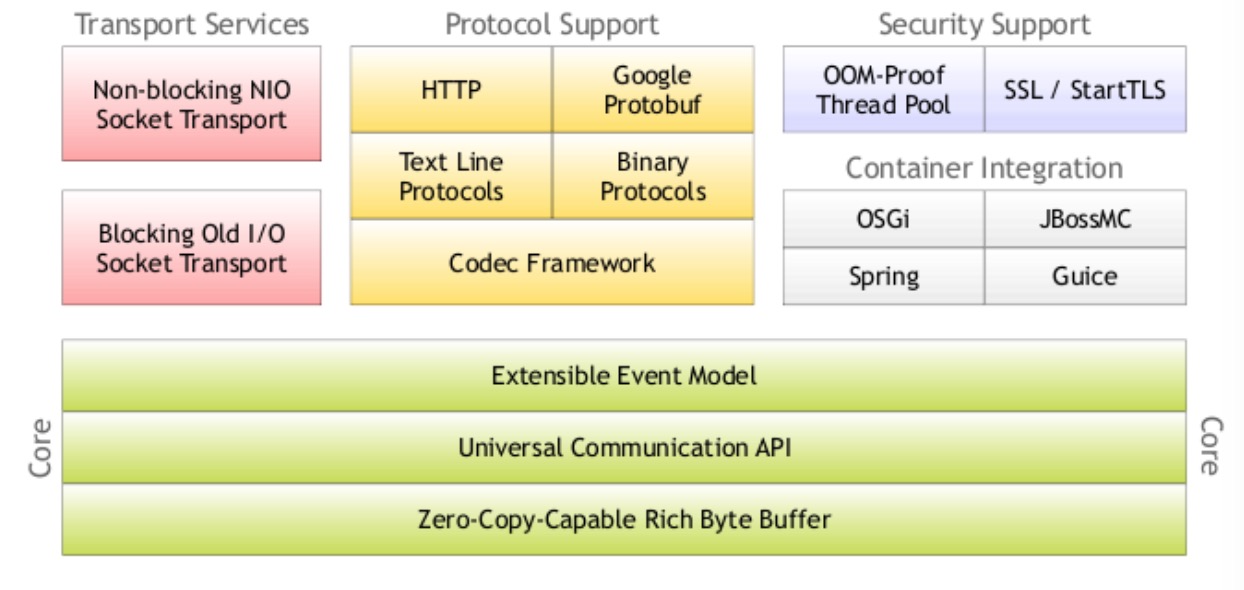
／／虚拟化，分布式存储，备份，压缩，安全，容错

第5章具体框架设计

4.1 netty框架的应用

4.1.1 netty 框架介绍

Netty是高效的Java NIO开发框架。它的总体结构如下：

从图中的最上层传输数据服务分为：阻塞模式和非阻塞模式。

阻塞模式基于TCP/IP协议，完整的服务包括服务端进程和客户端进程，在服务端申请ServerSocket进行，并注册监听端口，此时客户端可以建立Socket服务发起请求，自定对应服务的IP和端口，这时服务端可以通过ServerSocket的accept方法从请求队列中取出socket，如果此时没有建立好的链接，这时进程需要阻塞等待客户端的请求，对于已经建立好的链接，服务端和客户端可以通过输入输出流传送数据，对于这样的非阻塞模式一次只能处理一个请求，如果需要建立多个请求的时候，需要为每个请求分贝建立一个线程，这种模式可能会有一些线程处于等待状态造成资源浪费，线程数量越来越多，线程之间的切换也会带来越来越多的性能开销。

非阻塞模式是单线程操作，Selector用来实现非阻塞，其中的时间通知API可以确定一系列套接字中哪些已经处于就绪状态。由于在任意时间可以检查任意读操作和写操作的完成状态，这样就可以通过单个线程同时操作多个并发的链接，大大提高了内存的使用效率，使用较少的线程处理较多的请求，节省了上下文切换造成资源消耗，而且在高负载的情况下，可靠和高效地处理和调度IO操作不容易出错，这些复杂的处理都交给了Netty来处理。

现在Netty是一种采用基于NIO类库的异步通信框架，采用非阻塞模式，基于事件驱动，高性能，高可靠性的通信模式。

NIO采用Reactor模式，一个Reactor线程聚集了多路复用器Selector，它可以同时注册监听和轮训上百上千个Channel，一个线程就可以同时轮训处理多个客户端链接，线程的数量也可以是多个，但一般不会超过CPU个数+1，通过这种方式可以大大控制I/O线程之间上下文的切换和竞争，CPU的的利用率的到了很大的提高。所有I/O

操作是异步的，业务线程直接进行IO操作，也不会发生阻塞，系统不再依赖于外部的网络环境和外部程序的处理性能。这种非阻塞模式可以给系统带来很大的可靠性和性能提升。

Netty从逻辑上分为三层：

第一层是Reactor通信调度层，由一系列辅助类组成，包括Reactor线程NioEventLoop及其父类，NioSocketChannel／NioServerSocketChannel及其父类，ByteBuffer以及衍生出来的各种Buffer，Unsafe以及衍生出来的各种子类。

第二层是职责链ChannelPipeLine，主要负责上层调度时间在职责链中的传播，可以动态编排职责链，职责链可以拦截制定的时间，对其他的IO和事件忽略，Handler对inbound和outbound事件都支持。

第三层是业务逻辑编排层，这一层一般有单纯的业务逻辑编排和应用层协议插件，用于协议相关的编解码和链路管理。

Netty还可以实现数据在内存之间的零拷贝，即通常用户调用文件资源的时候，首先需要把文件拷贝到用户空间，而Netty可以直接在内核空间中传输到网络。Netty在进行接收和发送ByteBuffer使用的是直接buffers，使用非堆内存进行Socket的读写，不需要进行字节缓冲区的拷贝。如果使用传统的堆内存进行Socket的读写，就需要虚拟机首先把堆内存拷贝到直接内存中，然后才写入Socket中，这就多了一次缓冲区的内存拷贝。Netty使用了组合Buffer对象，可以聚合若干个ByteBuffer对象，用户可以像操作Buffer一样对组合Buffer进行操作，可以避免通过传统的方式将多个小的Buffer和并成一个大的Buffer。Netty允许我们将多段数据和行为一段虚拟数据供用户使用，这个过程也不会发生拷贝操作。Netty对数据的访问可以随机访问，根据index定位数据的位置，实现随机读写，同时为需要传输的数据拟定了统一的ChannelBuffer接口。

Netty的零拷贝是通过CompositeChannelBuffer来实现的，它的作用是将多个ChannelBuffer组成一个虚拟的ChannelBuffer来进行操作，这种虚拟的ChannelBuffer并没有将多个ChannelBuffer真正地合并起来，知识保存了数据的引用，这样就避免了数据拷贝；其中readerIndex和writeIndex是从AbstractChannelBuffer继承而来，components是一个ChannelBuffer类型的数组，它可以保存这个虚拟的Buffer的各个子Buffer，indices是一个int类型的数组，它保存的是各个Buffer的索引值，lastAccessedComponentId保存最后一次访问的子Buffer的ID，总的来说CompositeChannelBuffer就是一系列buffer通过数据结构包装起来，通过接口对外提供服务，让整体的Buffe组看起来就像是单个Buffer一样。

文件传输采用的是transterTo，它可以将文件缓冲区的数据发送到目标Channel中去，避免了传统的通过循环write方式导致的数据拷贝问题。

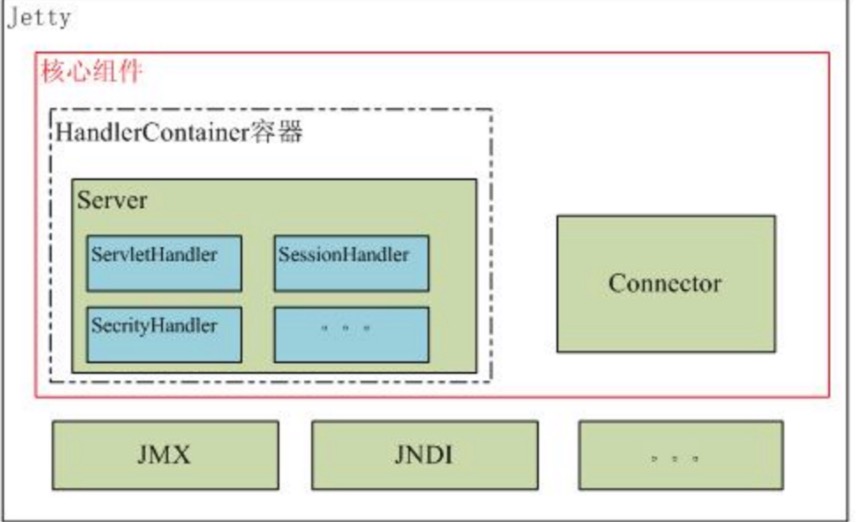
4.1.2 netty框架的应用

在我们的云存储中，Netty主要担任数据流读写的功能，客户发送的请求首先经过Cos-server的处理，将对应的数据切片，刷盘等交给IO-Server处理，这里包含了数据配片，高并发刷盘，消息队列，路由选择算法等，通过这一套服务，大大提高数据流的存储性能。对于每个大对象会进行分片，每个小的分片作为一个存储单位，并且存三份在不同的磁盘上，有效防止当机造成的数据丢失；对于设备的选择，使用了一套科学的路由选择算法，每次存储分片，会计算磁盘的存储经验值，选择相对来说读写性能较优的磁盘，例如读写较频繁的磁盘或者出现问题次数较多的磁盘，此时它对应的经验值较低，这样可以大大提高数据的操作性能。消息队列的使用可以提高客户的使用满意度，例如在删除数据的时候，拿到每个分片的URL和IP，生成任务放到MQ中，然后刷到Hbase的Task表中，这是就已经可以给客户返回操作成功的消息，实际的任务并没有做完，只是保存在了Hbase中，系统还有一个叫BaseRecover的进程，这个进程会定期扫描Task表，进行后续的任务处理。

4.2 jetty框架的应用

421 jetty框架的介绍

Jetty作为一个轻便的Web服务器，架构比较简单，有很强的扩展性，对于不同的业务可以灵活修改，Jetty有如下2个核心组件，Server组件是急于Handler容易工作，同时需要Connector的配合，负责在客户端发送请求是提供有效的连接，请求会被分配到一个处理队列中，当然也有一些其他的组件如JMX，JNDI等。



Jetty服务的主要是围绕着Server类来构建的，主要用来为Web程序提供访问服务，Jetty的Server扩展实现了一个个Handler，实现许多不同的逻辑处理。Jetty的Handler有2种类型，一种是HandlerWrapper，把Handler委托给另一个类去执行，如果把Handler加入到Jetty中去，这时我们就需要通过Server去调用，同时配合ScopeHandler类可以拦截Handler的执行，在Handler的前后可以添加一系列的操作，比如日志的打印，如果有数据库多步骤的操作，此时我们可以加进去事务，使程序的业务逻辑更加完善。另一种是HandlerCollection，这种方式是把多个Handler组装到一起，构成一个Handler链。

一个Jetty服务需要初始化Server服务，并注册web应用程序的Context，申明对应的路径，下面是一个最简单服务所需要的内容：

Server server = new Server(8080);

ServletContextHandler context = new ServletContextHandler(ServletContextHandler.SESSIONS);

context.setContextPath("/");

server.setHandler(context);

context.addServlet(new ServletHolder(new HelloServlet()), "/hello");

context.addServlet(new ServletHolder(new HelloServlet("Hello Kongxx!")), "/hello/kongxx");

context.addServlet(new ServletHolder(new GoodbyeServlet()), "/goodbye");

context.addServlet(new ServletHolder(new GoodbyeServlet("Goodbye kongxx!")), "/goodbye/kongxx");

server.start();

server.join();

在这里context可以添加任意多的业务处理Servlet，同时对于context我们也可以添加多个，保存各自的上下文：

Server server = new Server(8080);

ServletContextHandler context1 = new ServletContextHandler(ServletContextHandler.SESSIONS);

context1.setContextPath("/hello");

context1.setResourceBase(".");

context1.setClassLoader(Thread.currentThread().getContextClassLoader());

context1.addServlet(new ServletHolder(new HelloServlet("Hello Kongxx!")), "/kongxx");

ServletContextHandler context2 = new ServletContextHandler(ServletContextHandler.SESSIONS);

context2.setContextPath("/goodbye");

context2.setResourceBase(".");

context2.setClassLoader(Thread.currentThread().getContextClassLoader());

context2.addServlet(new ServletHolder(new GoodbyeServlet("Goodbye kongxx!")), "/kongxx");

ContextHandlerCollection contexts = new ContextHandlerCollection();

contexts.setHandlers(new Handler[] { context1, context2 });

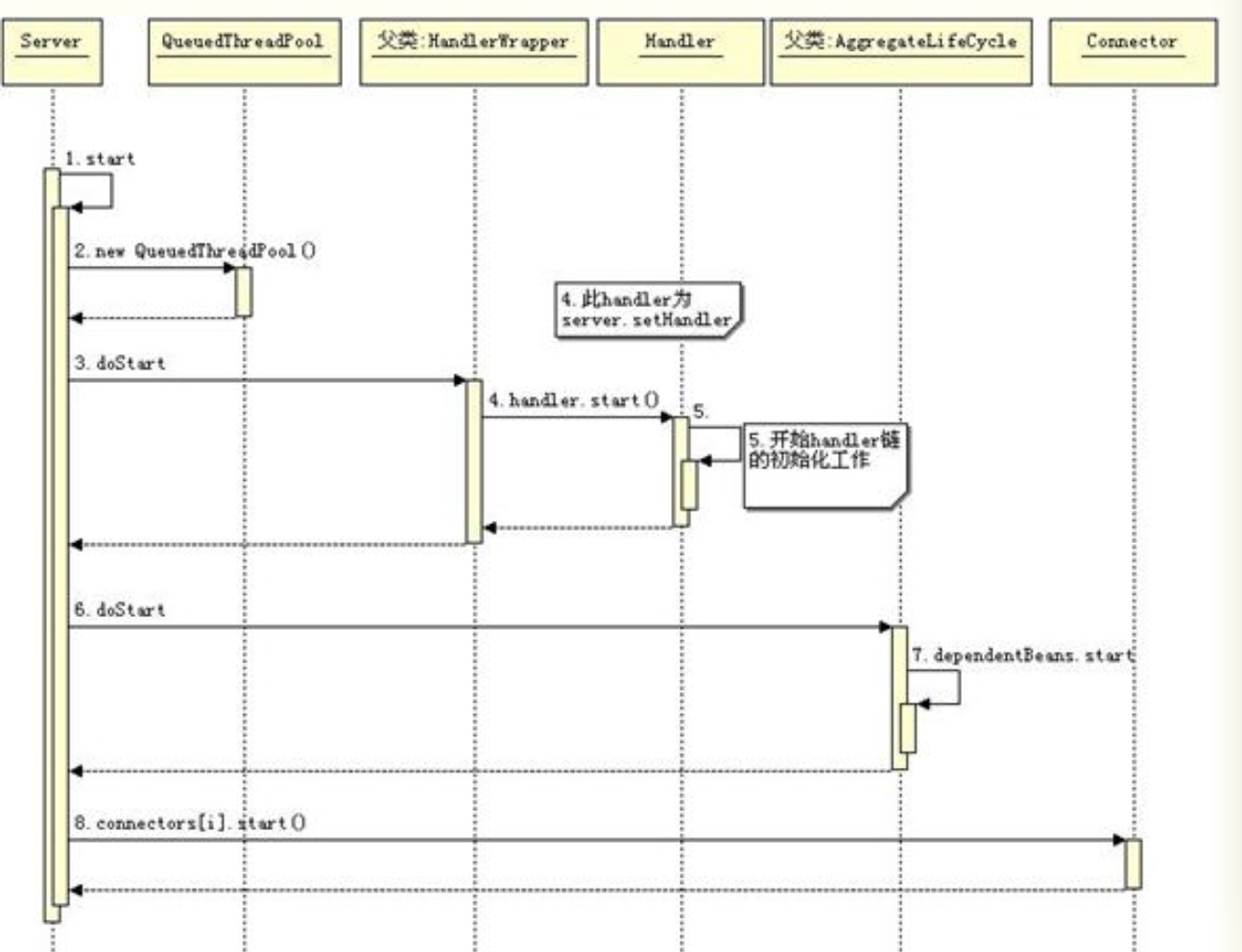
server.setHandler(contexts);

server.start();

server.join();

Jetty启动原理：

下图是启动以及服务调用的流程图：



Jetty服务的启动从Server的start函数开始，这个函数执行完后，代表Jetty能够对外提供服务了，至于需要哪些服务，可以对Jetty做不同的配置，Jetty中所有的组件都是继承自LifeCycle，在调用start函数的时候，会启动所有的已经注册的组件。首先Server会把注册的Handler启动，通常这个Handler会有很多的子Handler，这些子Handler形成Handler链，Server会依次把链上的所有Handler启动起来，然后让已经注册在Server上的JMX的Mbean启动启动起来配合Handler的相关工作，上述过程完成后会启动Connector，通过指定端口接受客户端的请求。

客户端和服务器的交互一般采用HTTP协议，当Jetty接收到客户发送过来的请求的时候，必须按照HTTP协议解析请求，从Header中获取相关的数据，进入到指定的类执行相应的业务逻辑，然后封装返回的数据。

处理客户端的链接实现类是org.eclipse.jetty.server.bi.SocketConnector，让Jetty以BIO的方式工作，BIO工作环境在启动的时候就会创建，并由HttpConnection 类来解析和封装Http协议；ConnectorEndPoint类是以BIO的方式处理链接请求，ServerSocket通过建立socket连接进行数据的交互；

Executor是处理connect的线程池，负责处理请求队列中的任务，acceptThread监听连接请求，下面是建立连接的过程：

首先创建一个队列线程池，用来处理每个建立连接产生的任务，线程池可以通过用户来指定；再建立ServerSocet，准备接受客户端的socket请求；最后再创建监听线程，这样可以用来监听访问端口是否有连接建立。

Acceptor线程会为客户端的请求创建ConnectorEndPoint，HttpConnection用来表示连接是一个http协议的连接，HttpPars类负责解析HTTP协议，同时创建Request和Response对象，然后把线程放到线程池中去执行。

基于AJP工作

一般来说，web网站的后台服务端不是将java程序的服务端直接暴露给用户的，比如说在Jboss的前面加一个web服务器，像apache或nginx都是常用的服务器，通过这种方式可以增加一些方便的功能如：日志分析，均衡负载，权限控制，防止外部的恶意请求以及一些静态资源的预加载等等。通过这种架构，servlet引擎就不再需要对http协议做处理，这些工作都在web服务器端完成了，极大地加快了请求的相应速度。

422jetty框架在云平台中的应用

对于整个云平台来说，jetty负责处理用户的第一层请求的，解析http协议，获取请求信息，web服务使用通过jetty和外部实现交互，这一层在云存储平台称之为cos-server，sdk可以通过裸写http请求申请云存储服务，也可以开发应用程序，调用cos-server的Api，所有用户需要实现的功能必须通过http请求通过cos-server向底层发送并获得服务，可以称之为云服务的包装层，整个云平台唯一可以跨越jetty服务层的是存储后台服务程序，这是云服务内部服务，不具有向其他外部开发者开放性，可以分配给若干个用户访问使用，程序功能较单一，没有扩展性，存储后台服务程序直接通过调用IO-server的Api调用调用相关的服务，这其实也是云平台对外提供存储服务的实例程序。

43Hbase框架

431Hbase框架原理

4311Hbase定义

Hbase是一个可靠，性能良好，列式存储的数据库，使用过程中可以根据需要动态扩大列的数量，通过Hbase可以通过大规模的廉价机器构建超大存储量的集群。Hbase的实现主要采用了Google Bigtable的原理，存储利用Hadoop HDFS作为文件存储系统，处理海量数据时使用Hadoop MapReduce来处理，集群中的海量机器通过zookeeper来管理。

4312 Hbase特性

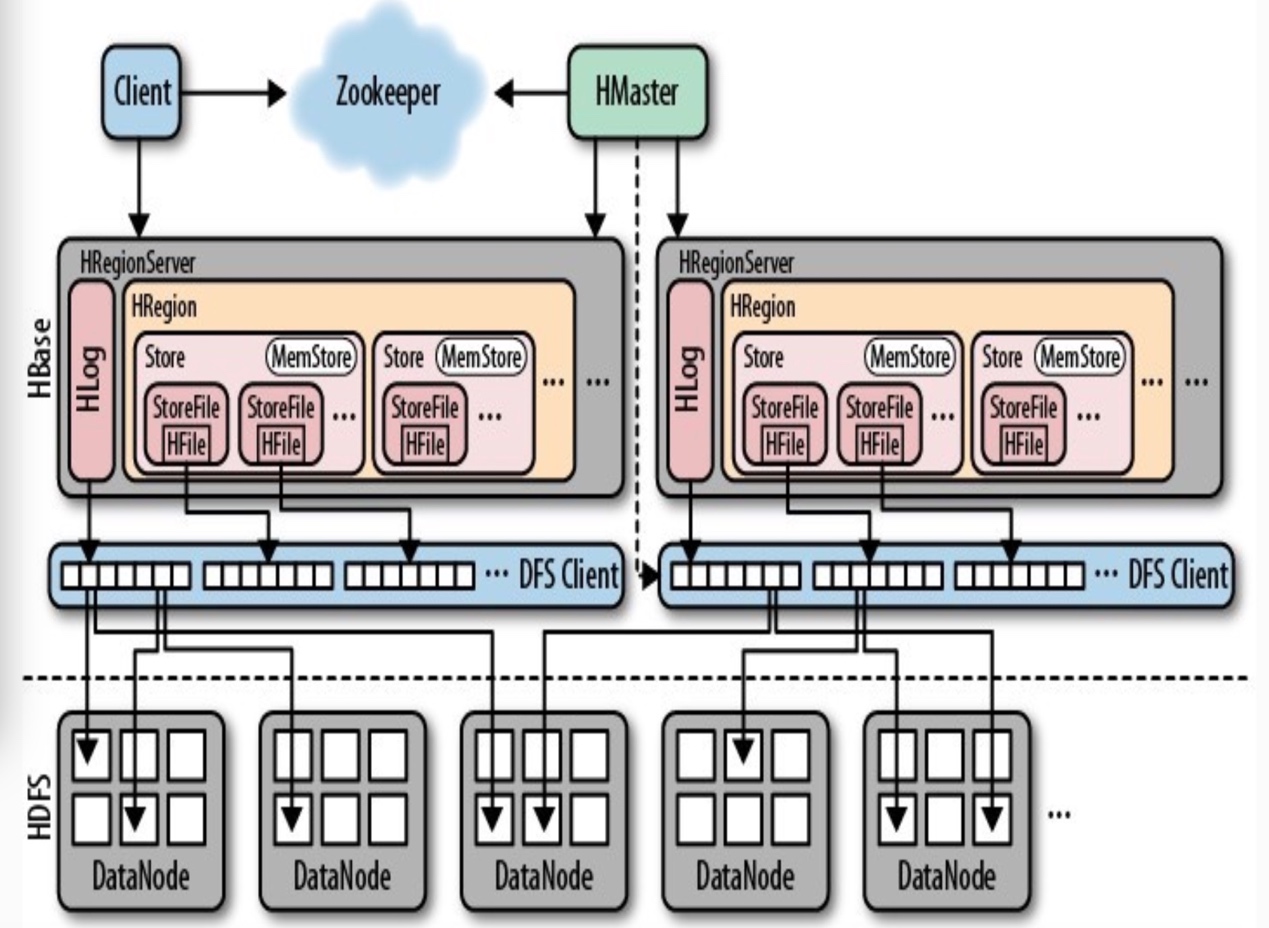
Hbase的表结构非常大，有上亿行和上百万列，且列可以动态增加，它的权限控制面对列簇，数据呈现列式块存储，列或者列簇检索信息独立，和传统的关系型数据库不同，空列不占用存储空间，这样表可以设计得非常稀疏。

4313Hbase访问接口

Hbase最常规和最有效的访问方式是Native Java API，适用于任务的并行处理；Hbase Shell ，命令行工具，一般来说不涉及复杂的业务，比如查询数据库所有的表，扫描表的数据，添加数据，增加列等等；Thrift Gateway，可以支持多种语言如C++，JS等的访问；REST Gateway，支持Http的API访问Hbase服务，不受客户端语言的限制，在Hbase服务端解析相应的Http请求即可；Pig可以使用Pig Latin流失编程语言调用库里的数据，它的原理就是通过程序生成MapReduce Job来处理Hbase的表数据，对海量数据的处理有很大的优势；Hive，通过编写类似的SQL语句来访问库里的数据。

4314 Hbase 的结构

下图是Hbase的存储结构模型图：



Hbase的主要包含Hmaster，HregionServer，HRegion，Store，MemStore，StoreFile，Hfile，Hlog等模块，下面将对各个模块作详细的介绍。

Hmaster： 可以为Region server分配region，资源使用不均衡是，调整不同的server之间的负载均衡，当有的设备出现故障，此时可以将此server上的region重新分配到别的设备上，随着业务额扩展，文件系统中会逐渐初心啊一些垃圾文件，Hmaster会负责这部分文件的回收，释放出来这部分存储资源，对于一些schema的请求也是H master处理的。

HRegionServer：主要管理被分配的Region，处理在其上的IO请求，随着存储数据量的增加，Region逐渐扩大，需要把Region切分，这是HregionServer的一个很重要的功能。通常用户在访问Hbase中表的数据时，并不需要master作处理，对于客户的请求，首先需要访问zookeeper和region server，然后访问数据通过region server，master只需要处理好表和region的元数据信息即可，负载比较低。HregionServer存取一个表时，会创建一个HRegion对象，表的列簇是查询的单位，每个列簇对应着一个Store实例，这其中有一个MenStore和若干个StoreFile与之对应，每个StoreFile又对应一个HFile，这里的HFile是存储在磁盘中的文件。

Hregion：一个表上有有若干个Region，Region是分布式存储的最小单位，Region可以放在不同的设备上，一个Region必须存储在同一个设备上，对于Region来说，Hbase会给Region的大小设置一个阀值，随着数据存储量的增加，Region包含的信息越来越多，当超过了这个阀值，此时就会被分成2个新的region。Hbase定位region的过程为：首先通过zookeeper的存储region位置的文件找到-ROOT-表的位置，这里-ROOT-表是.META.的第一个Region，而.META.表中的每一个region在-ROOT-表都有一条记录，这时可以通过.META.表找到所要的用户表region的位置。

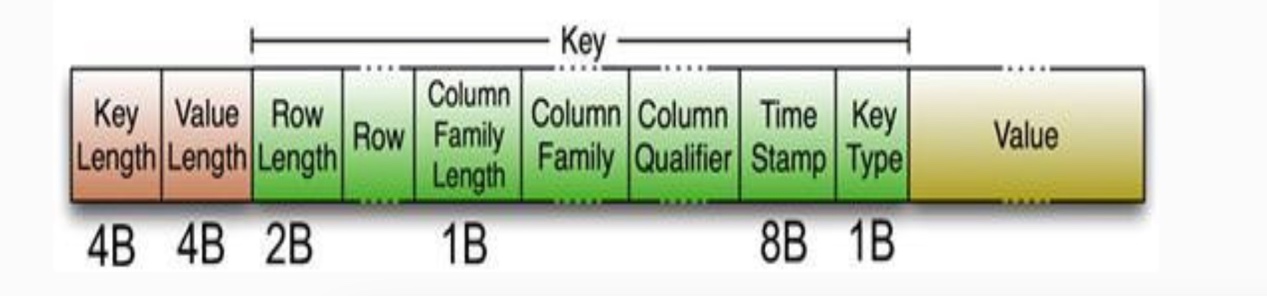
一般来说一个表开始只有一个region，随着数据量的增加，region会进行分裂，每个region都有startkey和endkey参数。

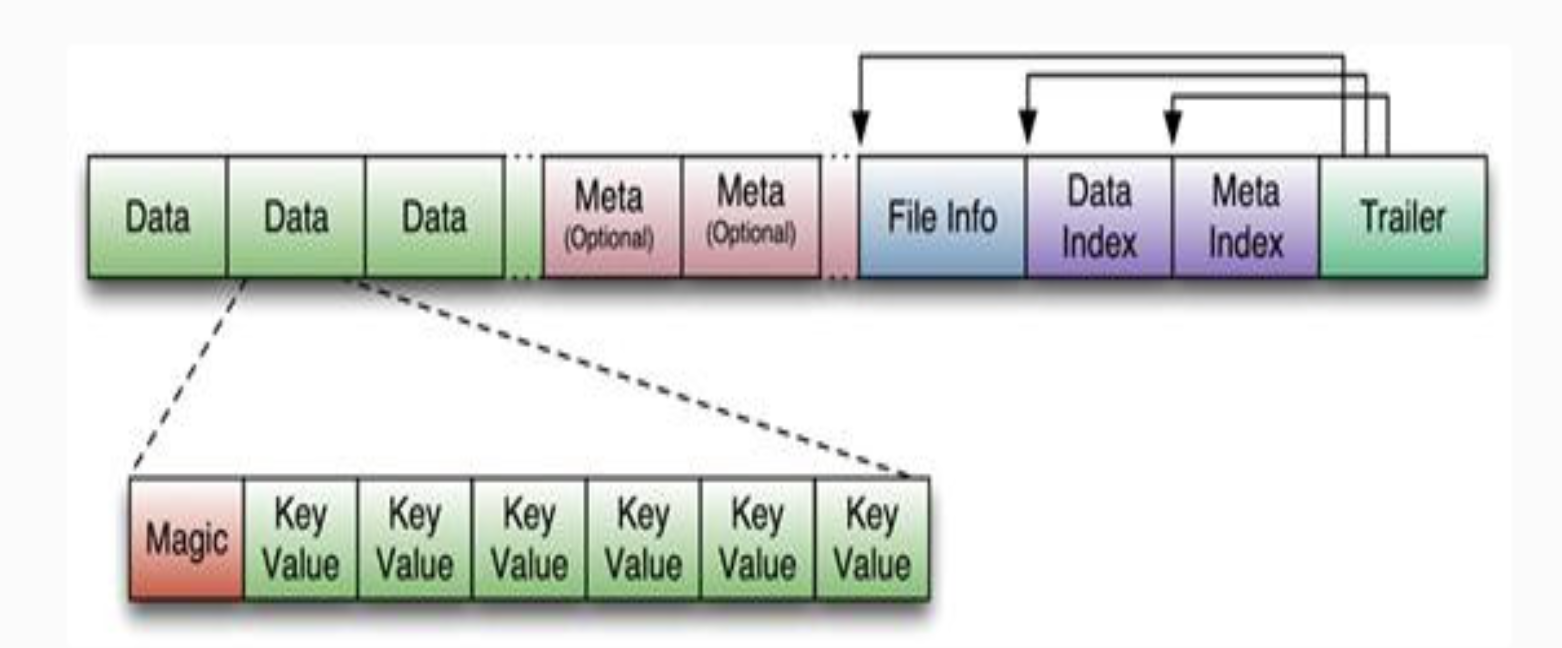
Store：region主要是由若干个store组成的，每个列簇建一个store，而一个store由一个memStore和0或者多个StoreFile组成。HBase以store的大小来判断是否需要切分region。

MemStore：memstore放在内存里，数据以keyValues的形式，由于存储在HDFS上的数据需要按照行键进行排序，HDFS本身就是顺序读写，不能修改，这样的话Hbase不能高效写数据，写到Hbase的数据是不会被排序的，将来检索性能大大降低了，于是Hbase为了解决这个问题，就在内存中多了一步处理，完成数据的排序，这样写入到磁盘的数据是顺序的，为将来的数据检索带来了性能上的提升。

StoreFile：memStore内存中的数据写到文件后就是StoreFile，在底层以HFile的形式存在。

HFile：在Hbase中，数据主要以键值对的形式存储，是二进制文件。HFile的文件长度长短不一，其中有2部分是固定的长度：Trailer和FileInfo，Trailer中含有其他的数据块的起始位置信息，FileInfo记录了文件的一些元数据信息，数据块是Hbase进行输入输出的最小单位，对于数据块的大小可以在建表的时候指定，如果不指定，一般来说，默认大小是64KB，块的大小各有利弊，较大的块方便顺序扫描，小的块有利于随机查询。每个块由一个Magic和多个键值对组成，Magic储存一些随机化的数字来保证数据的完整性的，每个KeyValue包含了键长，值的长度，行的长度，列簇长度，列的长度，列属性，时间戳以及key类型等。数据块主要是用来保存表中的数据的，数据块可以压缩。元数据段用来保存一些自定义的键值对，这部分可以压缩，而文件信息段主要用来保存HFile的元信息的，则不可以压缩，data index用来保存meta块的索引，Trailer主要是用来保存每一段的偏移量，读取HFile的时候，首先会读取Trailer，这样就找到了data index，进而不需要遍历整个文件就能找到对应的数据块，再找出最终的Value。现在的存储模式，数据块和元数据块基本都采用的压缩的方式，压缩可以大大减少IO的时间，不过这样也又一个弊端就是压缩和解压耗费了一定的资源，常用的压缩方式有gzip和lzo。





Hlog：Hlog文件就是一个普通的序列化文件，序化化文件的value的key是HlogKey对象，记录了写入数据的归属信息，存储的信息包括了table名，region名，时间戳（写入时间），对应的value是keyvalue对象，也就是HFile中的键值对。Hlog主要是数据写入钱，先把日志保存下来，当机器发生故障的时候，可以从Hlog中调出数据变更的记录，可以通过这些记录进行恢复。

LogFlusher：存储数据的时候，首先把日志写下，然后再写入序列文件，通常来说，数据流进入文件系统，采用缓存机制，保证存储的性能，可能有时一部分数据对于客户端来说已经显示存储完成，实际上，这部分的数据在服务端并没有真正刷到磁盘，而是存在内存中。这就需要处理类LogFlusher，通过调用optionalSync()方法，在hbase还有一个设置默认定时刷日志的设置项hbase.regionserver.optionallogflushinterval，一般情况下默认是10s，这个参数定期调用Hlog.sync()，进行数据刷盘；还有一个doWrite函数也会定期调用sync()，实现刷盘功能。

LogRoller：Log的大小通过配置文件hbase.regionserver.logroll.period来设置，不设置的话，默认是一个小时，每过60分钟，便会产生一个新的log文件。这样随着时间的推移，会形成很多很多的log文件需要定期整理，rollerWriter定期滚动日志，有一些日志也可以定期清除。每个region server都会维护一个Hlog，这样，不同的region的日志就混杂在了一起，这样在追加单个文件的时候，可以同时写多个文件，大大减少了磁盘的寻址次数，table的写性能得到了很大的提高。这样也有相应的弊端，如果某台设备出现故障了，对数据进行恢复时，需要把日志进行拆分，然后发送到其他的region server进行恢复。

Hbase 结构设计

Hbase的数据库表是一张很大的表，从整体上看，行记录有三个基本类型，包括：RowKey，Time Stamp，Column。RowKey是数据在表中的唯一标识，时间戳是数据版本的标志，列由列簇和列名组成，列簇的修改类似与需要修改表结构，而列名可以动态增加，这也是Hbase能够动态增加列的体现；列示存储使得读写性能得到很大的优化，相同的family的数据存储位置比较近，对于这样的存储特性，在我们处理具体访问需求时，我们可以通过业务逻辑层加以利用。

在Hbase数据库中，行键用来检索数据的，我们可以通过单个row key进行访问，全表扫描访问，也可以使用行键的前缀进行访问，在数据底层，row key保存为字节数组，而存储的时，按照字典顺序进行存储，设计Key的时候，我们需要充分利用这个特性，把经常一起读写的数据放在同一个列簇中。

列簇包含若干个表，列簇是Schema的一部分，在定义表时就需要指定好，而列就不需要，是可以动态调整的属性，每个列都有其所属的列簇，对于访问的控制，磁盘和内存的使用统计都在列簇的层面上的，访问控制可以允许一些应用有修改数据的权限，或者只有浏览数据的权限等等。

通过行和列簇可以确定存储单元。Cell的数据没有具体的类型，全部以字节码形式存储，每个Cell都保存着多个版本，一些数据会进行覆盖和内容的修改，进行数据变动时，系统会指定此时的时间戳，也可以自行添加时间，Cell的不同的版本倒序排列，最新的数据在前，一般访问数据时，只需要最新的那一版数据即可，对于数据的管理版本，目前有2总常用的数据管理方法，一种是按照最近的时间去管理，比如根据存储能力的大小，数据版本更新的频率，数据量的大小等，规定一个可以存在多版本数据可控范围的时间区间，比如一月，这样，所有一个月内的数据增加和修改版本都会保存下来；另一种方法是规定版本的数量，每条数据指保存N个最近的版本，增加一个新的版本，如果此时已经达到存储的数量上限，此时需要存储最新的版本，删除最老的那一版数据。

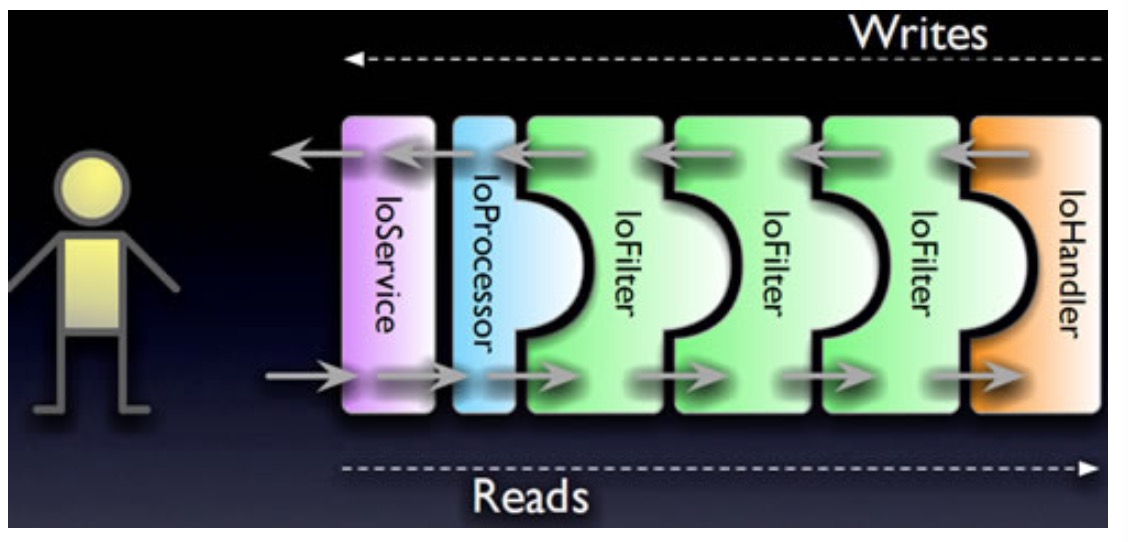
432 Hbase的应用

作为大型的云存储平台，对数据的扩展性和密集访问性有着很高的要求，高峰时期，可能有每秒上千次的访问量，对这样的高频访问，对于普通的关系型数据库来所，不可能完成，得益于Hbase的行键访问机制，能够迅速从海量数据中挑选出我们想要的数据，所以选择Hbase作为云平台的元数据存储，Hbase依赖于HDFS，本平台的流数据存储在HDFS上，这样也能通过应用程序获得元数据的同时，快速获得真实数据。目前Hbase已经开源，随着功能和性能的完善,目前Hbase的性能也相对稳定，具有很强的可扩展性，数据库表对应实现的实例类，很方便实现了对数据库的操作。对于数据的版本，我们可以保存多份最近修改的数据，方便对历史版本的调用。

对于Hbase的行键搜索是通过顺序搜索的，这样对于大对象的存储，采用分片存储的功能，不同的分片采用相同的Hash值加对应的序号，对于大对象进行切割，在存储数据时，可以实现多个分片同时上传到不同的磁盘上，在检索的时候，通过对应的Hash值模糊查询出所有的分片，然后组合成完整的对象，再传给客户端。对于像批量删除大批数据这样的操作，采用在Hbase中建各种task表，失败表，未处理表，处理失败的表，这些表用来暂时存储客户发过来的任务，可以在人物还没有真正完成的时候，给客户一个提前的反馈，然后系统可以选择一个合适的时间去执行这个任务，并在对应的位置，如bucket上加相应的标志表明某文件已经被删除，这样在刷新bucket的时候，就不会再把不需要的文件也罗列出来。在Hbase中还设计了流量监控表，对不同的用户访问情况作相应的统计，对后期的收费问题，需要一些参数比如某个用户的访问次数，上传次数，下载次数，流量大小等等，这些都是消费大小的指标，这些服务功能都通过Hbase高效管理地了起来。

44 MINA框架

441MINA框架介绍



Apache Mina Server是一个应用与网络通信的框架，底层通信协议可以是TCP/IP,UDP/IP协议，当我们的应用程序需要较高的处理性能和较高的扩展性，MINA框架是一个不错的选择，MINA提供了事件驱动操作，底层使用NIO机制，刚刚发行是1.x版本，现在我们所说的MINA框架一般都是指2.x版本，框架基本已经把底层的通信机制和业务逻辑隔离开来了，作为程序开发人员来说，只需要关注发送的数据和业务逻辑就可以了，下面将对框架的各层作详细的讲解：

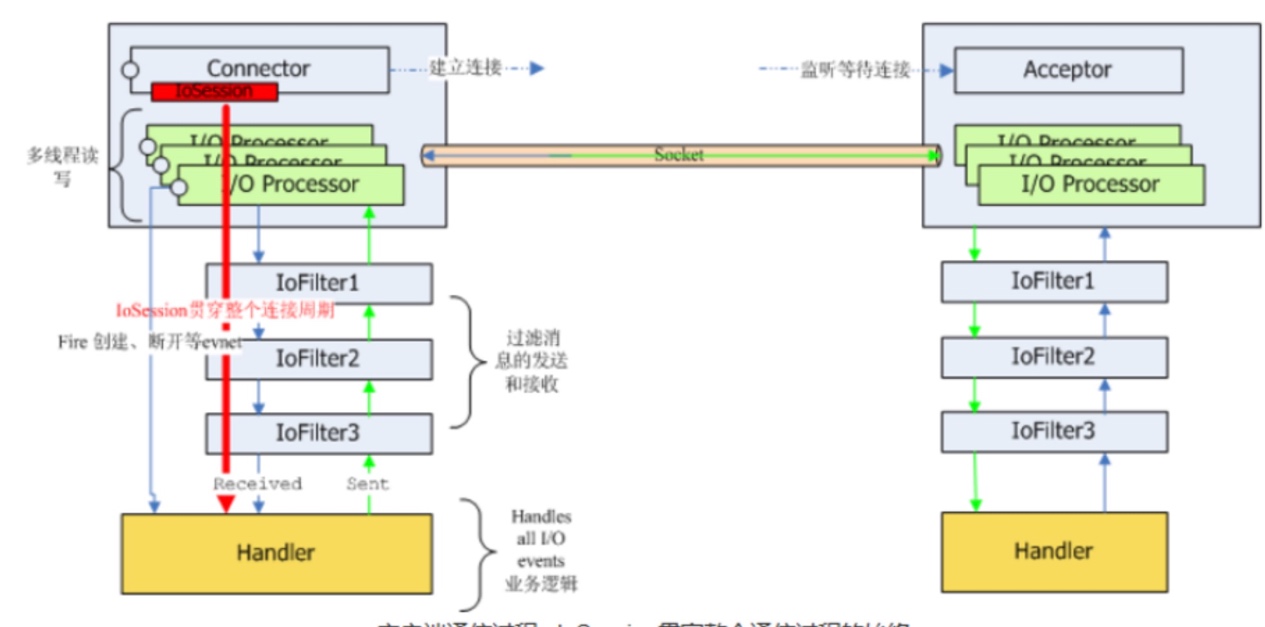
IoService: 这是框架的最底层，主要用来处理IO相关的工作的。IOSocketAcceptor对应TCP协议下面的服务端IOService，而IOSocketCHannel对应TCP协议下的客户端IOService，IOService主要用于隐藏底层IO操作的细节，对上层提供统一的基于事件的异步IO接口。数据到达时，IOService调用底层IO接口读取数据，然后封装成IoBuffer，然后以事件的形式传到上层，这样底层的同步IO转化成了异步IO，low level IO也经过IOService变成了IO Event。

IoProcessor：系统需要为此接口单独起一个线程，负责检查是否有数据在通道上读写，这样它就拥有了自己的Selector，这与通常的NiO编码不一样，一般来说只有一个Selector， IoProcessor负责调用注册在IoService上的过滤器，然后调用IOHandler。

IoFilter：这个借口定义一组拦截器，比如说一些日志的输出，系统设置的一些限制用户，对数据进行编码和数据解码等相关功能，在开发过程中，很多时候会遇到数据的编码解码问题，这对数据的传输有着至关重要的作用，在内存中和网络中，数据的传输完全不一样，所以一个合格的开发人员需要考虑的不仅仅是功能的实现，更加需要关注程序运行的效率，这就尤其需要我们关注数据的编码。

IoHandler：这一层是开人员需要重点开发的一层，也是数据接收发送的层，在这里我们会对发送过来的资源进行相关的处理，这一层的具体实现需要开发人员作详细设计和处理，这一层是框架处理流程的最后一站，通畅在这一层会对业务做详细的包装，对于一个指令的执行，做一系列周到的判断，指令执行的条件是否满足，调用相关的类方法，在本流程业务处理过程中，对于不合理的操作，或者操作条件不满足，会有一系列异常机制，并将异常信息反馈给客户端，对于这一层有完善的继承实现机制，严格按照框架指定的继承方式去实现相关的功能，每个IoService都需要指定一个IoHandler。

IoSession：这个类是对底层连接的封装，当新的连接建立的时候，会立刻生成一个context和session，用来存储全局信息，保证这些信息在执行的过程中能够随之拿出这些信息作相关的应用，发送数据是一个异步的过程，业务层首先会逆向穿过Filter链，到达IoService，此时并没有直接调用底层的接口实现相关的数据操作，而是将这次调用封装成一个写请求，然后放到session的请求队列中去，通过IoProcessor线程统一flush出去，这样就不会造成线程的等待情况。



442MINA框架应用

45 HDFS

第5章云存储服务平台的详细实现

51COS-SERVER

52 WEB-SERVER

53 SDK

54 FTP

第6章 总结

6.1 总结

／／遇到的问题，开发的难度

6.2 展望

附录A

HBASE数据库

参考文献

致谢