厦门大学计算机科学系研究生课程

《大数据技术基础》

第5 章 HDFS

（2013年新版）

林子雨

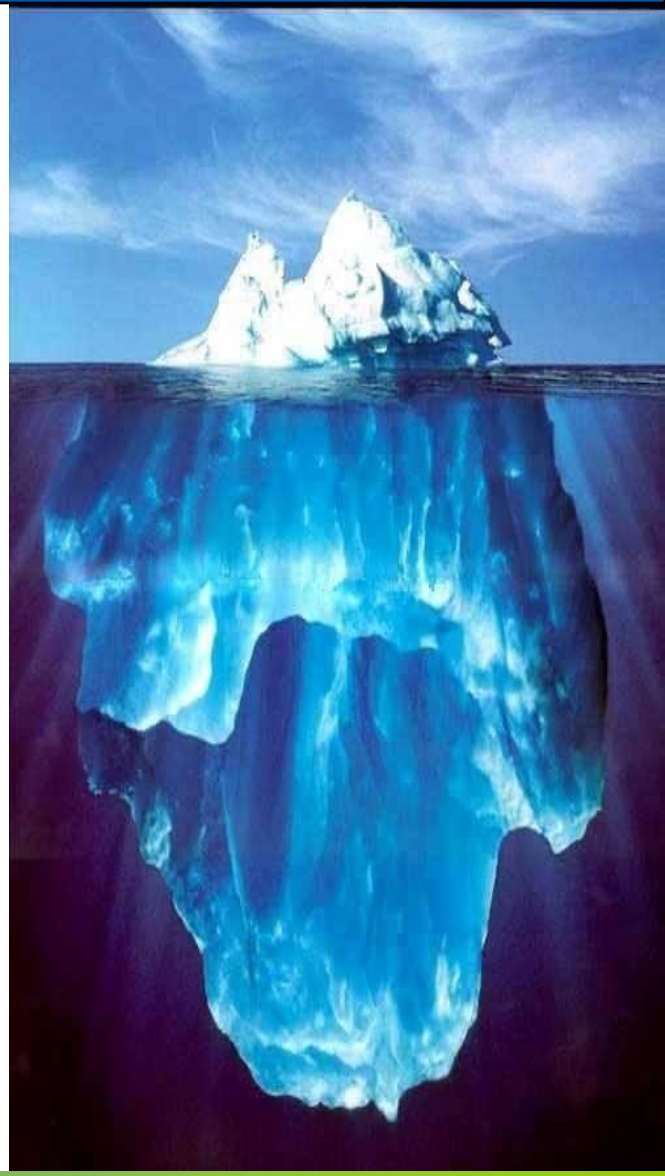
厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

提纲

- **HDFS的假设与目标**
- **HDFS的相关概念**
- **HDFS体系结构**
- **HDFS命名空间**
- **HDFS存储原理**
- **通讯协议**
- **数据错误与异常**
- **从HDFS看分布式文件系统的设计需求**



本讲义PPT存在配套教材，由林子雨通过大量阅读、收集、整理各种资料后编写而成
下载配套教材请访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423

HDFS的假设与目标

HDFS是基于流数据模式访问和处理超大文件的需求而开发的，它可以运行于廉价的商用服务器上。HDFS在设计时的假设和目标包括以下几个方面：

- 硬件出错
- 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

正是由于以上的种种考虑，我们会发现，现在的HDFS在处理一些特定问题时，不但没有优势，而且有一定的局限性，主要表现在以下几个方面：

- 不适合低延迟数据访问
- 无法高效存储大量小文件
- 不支持多用户写入及任意修改文件

HDFS的相关概念——块

•HDFS的块是抽象的概念，它比操作系统中所说的块要大得多。在配置Hadoop系统时会看到，它的默认块为64MB。和单机上的文件系统相同，HDFS分布式文件系统中的文件也被分成块进行存储，它是文件存储处理的逻辑单元。

•HDFS作为一个分布式文件系统，是设计用来处理大文件的，使用抽象的块可以带来很多好处

- 可以存储任意大的文件，而又不会受到网络中，任一单个节点磁盘大小的限制
- 使用抽象块作为操作的单元，可以简化存储子系统
- 块更有利于分布式文件系统中复制容错的实现

HDFS的相关概念——NameNode和DataNode

HDFS体系结构中有两类节点，一类是NameNode，另一类是DataNode。这两类节点分别承担Master和Worker的任务。

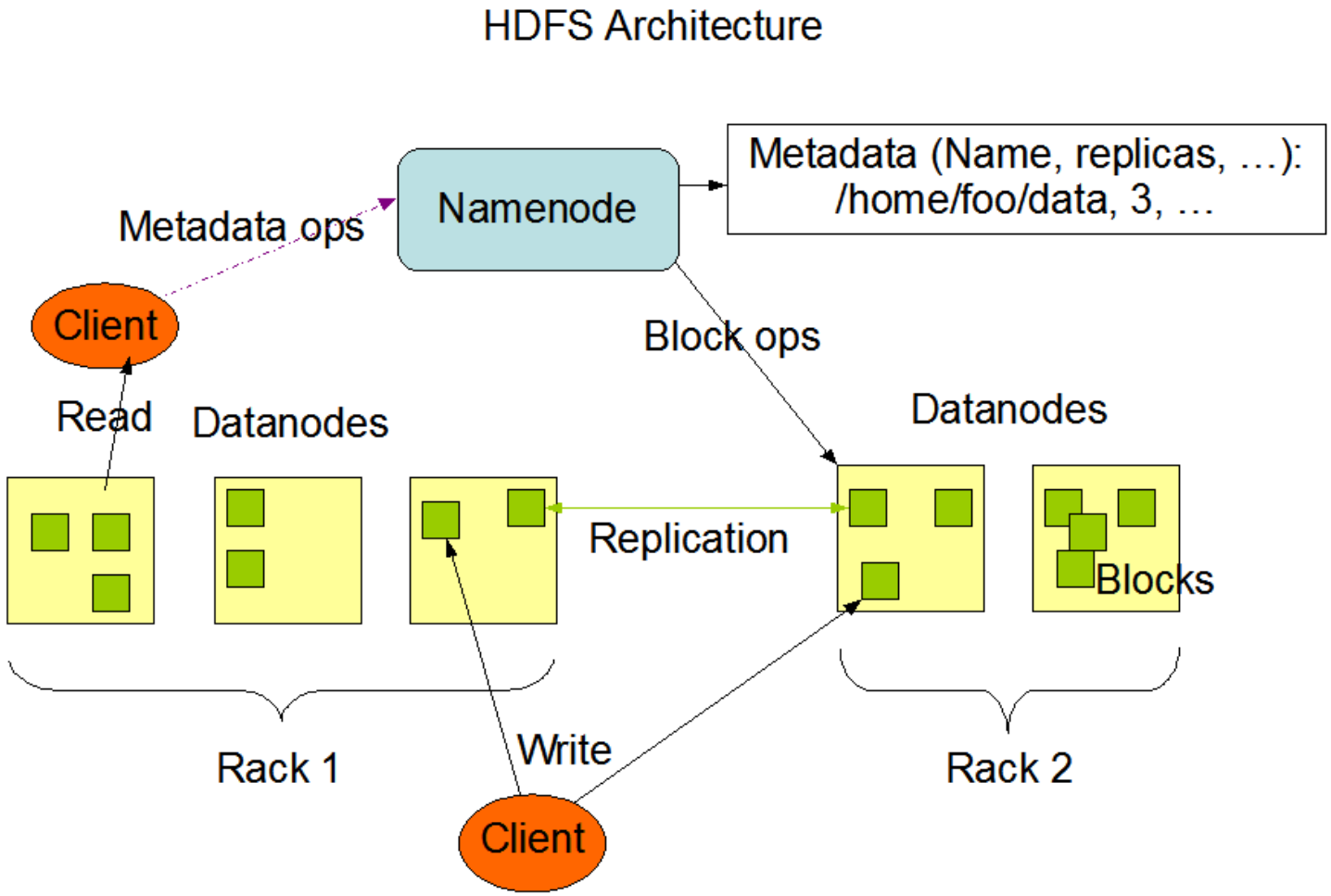
目录节点（NameNode）是集群里面的主节点，负责文件名的维护管理，也是客户端访问文件 的入口。文件名的维护包括文件和目录的创建、删除、重命名等。同时也管理数据块和数据节点的映射关系，客户端需要访问目录节点才能知道一个文件的所有数据块都保存在哪些数据节点上。

数据节点（DataNode）一般就是集群里面的一台机器，负责数据的存储和读取。在写入时，由目录节点分配数据块的保存，然后客户端直接写到对应的数据节点。在读取时，当客户端从目录节点获得数据块的映射关系后，就会直接到对应的数据节点读取数据。数据节点也要根据目录节点的命令创建、删除数据块和冗余复制。

HDFS体系结构

- Hadoop文件系统采用主从架构对文件系统进行管理，一个HDFS集群由唯一一个目录节点（NameNode）和数个数据节点（DataNodes）组成。
- HDFS对外表现为一个普通的文件系统，用户可以用文件名去存储和访问文件，而实际上文件是被分成不同的数据块，这些数据块就是存储在数据节点上面。
- 一个典型的Hadoop文件系统集群部署，是由一台性能较好的机器运行目录节点，而集群里面的其它机器每台上面运行一个数据节点。
- 唯一的目录节点的设计大大简化了整个体系结构，目录节点负责Hadoop文件系统里面所有元数据的仲裁和存储。这样的设计使数据不会脱离目录节点的控制。

HDFS体系结构


HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

HDFS命名空间

- Hadoop文件系统使用的是传统的分级文件体系，客户端程序可以创建目录并且在目录里面保存文件，类似于现在一般的文件系统。Hadoop允许用户创建、删除文件，在目录间转移文件，重命名文件等。

- 目录节点负责存储和管理整个文件系统的命名空间，应用程序可以指定某一个文件需要在Hadoop文件系统中冗余多少份，这个在Hadoop中称为冗余因子，保存在目录节点里面。

HDFS存储原理——冗余数据保存

•Hadoop文件系统是为了大文件的可靠保存而设计的，一个文件被划分成一连串的数据块，除了文件的最后一块以外其它所有的数据块都是固定大小的，为了数据容错性，每一个数据块都会被冗余存储起来

•目录节点是根据数据块的冗余状况来作出处理决策的，数据节点会定期发送一个存在信号（Heartbeat）和数据块列表给目录节点，存在信号使目录节点认为该数据节点还是有效的，而数据块列表包括了该数据节点上面的所有数据块编号

HDFS存储原理——数据存取策略

复制策略是Hadoop文件系统最核心的部分，对读写性能影响很大，Hadoop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，这个特性需要很多时间去优化和调整。

□数据存放

目前hadoop采用以机柜为基础的数据存放策略，这样做的目的是提高数据可靠性和充分利用网络带宽。

一个大的Hadoop集群经常横跨多个机柜，而不同机柜之间的数据通讯同经过交换机或者路由，所以同一个机柜中不同机器的通讯带宽是比不同机柜之间机器通讯时候的大。

Hadoop提供了一个API来决定数据机所属的机柜ID，当文件系统启动的时候，数据机就把自己所属的机柜ID发给目录机，然后目录机管理这些分组。

HDFS存储原理——数据存取策略

Hadoop默认是每个数据机都是在不同的机柜上面，这种方法没有做任何性能优化，但是也有不少**优点**：

- 数据可靠性是最高的
- 在读取数据的时候充分利用不同机柜之间的带宽
- 可以很容易完成负载平衡和错误处理

缺点:写入数据的时候并不能完全利用同一机柜里面机器的带宽。

在默认的配置下，Hadoop的冗余复制因子是3，意思就是每一块文件数据一共有3个地方存放，Hadoop目前的存放策略是其中两份放在同一个rack id的不同机器上面，另外一个放在不同rack id的机器上面，简单来说就是1/3的冗余数据在一个机柜里面，2/3的冗余数据在另外一个机柜里面，这样既可以防止机柜异常时候的数据恢复，又可以提高读写性能。

HDFS存储原理——数据存取策略

□数据读取

数据读取策略是：根据前面所说的数据存放策略，数据读取的时候，客户端也有api确定自己的机柜id，读取的时候，如果有块数据和客户端的机柜id一样，就优先选择该数据节点，客户端直接和数据节点建立连接，读取数据。如果没有，就随机选取一个数据节点。

□数据复制

数据复制主要是在数据写入和数据恢复的时候发生，数据复制是使用流水线复制的策略。当客户端要在Hadoop上面写一个文件：

- 首先它把这个文件写在本地
- 然后对文件进行分块，默认64MB一块
- 每块数据都对Hadoop目录服务器发起写入请求
- 目录服务器选择一个数据机列表，返回给客户端
- 然后客户端就把数据写入第一台数据机，并且把列表传给数据机
- 当数据机接收到4KB数据的时候，写入本地，并且发起连接到下一台数据机，把这个4K传过去，形成一条流水线。
- 当最后文件写完的时候，数据复制也同时完成，这个就是流水线处理的优势。

HDFS通讯协议

Hadoop的通讯协议基本是在TCP/IP的基础上开发的，客户端使用ClientProtocol和目录服务器通讯，数据机使用DatanodeProtocol和目录服务器通讯，而目录服务器一般只是应答客户端和数据机的请求，不会主动发起通讯。

数据错误与异常

Hadoop文件系统的主要目标就是在硬件出错的时候保证数据的完整性，它把磁盘错误作为肯定会出现的情况来对待，而不是异常。一般数据存储中出现的错误有几种，分别是目录服务器错误，数据机错误，和网络传输异常。

1.数据机出错

每个数据机会定时发送一个心跳信息给目录服务器，表明自己仍然存活，网络异常可能会导致一部分数据机无法和目录服务器通讯，这时候目录服务器收不到心跳信息，就认为这个数据机已经死机，从有效I/O列表中清除，而该数据机上面的所有数据块也会标记为不可读。这个时候某些数据块的冗余份数有可能就低于它的冗余因子了，目录服务器会定期检查每一个数据块，看看它是否需要进行数据冗余复制。

数据错误与异常

2.出现数据异常

由于网络传输和磁盘出错的原因，从数据机读取的数据有可能出现异常，客户端实现对数据块的校验，用md5和sha1进行校验，客户端在创建文件的时候，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机读取该文件块，并且报告给目录服务器这个文件块有错误，目录服务器就会定期检查，并且重新复制这个块。

3.目录服务器出错

FsImage和Editlog是目录服务器上面两个最核心的数据结构，如果其中一个文件出错的话，会造成目录服务器不起作用，由于这两个文件如此重要，所以目录服务器上面可以设置多个备份文件和辅助服务器，当这两个文件有改变的时候，目录服务器就会发起同步操作，虽然这样增加了系统的负担，但是在目前这个架构上面为了实现数据的可靠性，这个同步操作是非常必要的。

从HDFS看分布式文件系统的设计需求

分布式文件系统的设计目标大概包括：透明性、并发控制、可伸缩性、容错以及安全需求等。从这几个角度去观察HDFS的设计和实现，可以更清楚地看出HDFS的应用场景和设计理念。

- 透明性
- 并发控制
- 文件复制功能
- 硬件和操作系统的异构性
- 容错能力
- 安全性问题

从HDFS看分布式文件系统的设计需求

- 透明性

如果按照开放分布式处理的标准确定就有8种透明性：访问的透明性、位置的透明性、并发透明性、复制透明性、故障透明性、移动透明性、性能透明性和伸缩透明性。对于分布式文件系统，最重要的是希望能达到5个透明性要求：

（1）访问的透明性
（2）位置的透明性
（3）移动的透明性
（4）性能的透明性和伸缩的透明性

从HDFS看分布式文件系统的设计需求

并发控制

•客户端对于文件的读写不应该影响其他客户端对同一个文件的读写。
•HDFS的机制非常简单，任何时间都只允许一个写的客户端，文件经创建并写入之后不再改变，它的模型是一次写，多次读。
•这与它的应用场合是一致的，HDFS的文件大小通常是MB至TB级的，这些数据不会经常修改，最经常的是被顺序读并处理，随机读很少，因此 HDFS非常适合MapReduce框架或者web crawler应用。
•HDFS文件的大小也决定了它的客户端不能像某些分布式文件系统那样缓存常用到的几百个文件。

从HDFS看分布式文件系统的设计需求

文件复制功能

一个文件可以表示为其内容在不同位置的多个拷贝，这样做带来了两个好处：（1）访问同一个文件时，可以从多个服务器中获取，从而改善服务的伸缩性；（2）提高了容错能力，某个副本损坏了，仍然可以从其他服务器节点获取该文件。HDFS文件的块，为了容错都将被备份，并且可以配置复制因子，默认是3。副本的存放策略也是很有讲究的，一个放在本地机架的节点，另一个放在同一机架的另一节点，还有一个放在其他机架上。这样可以最大限度地防止因故障导致的副本丢失。不仅如此，HDFS读文件的时候也将优先选择从同一机架乃至同一数据中心的节点上读取块。

硬件和操作系统的异构性

由于构建在Java平台上，HDFS的跨平台能力毋庸置疑，得益于Java平台已经封装好的文件IO系统，HDFS可以在不同的操作系统和计算机上实现同样的客户端和服务端程序。

从HDFS看分布式文件系统的设计需求

容错能力

在分布式文件系统中，尽量保证文件服务在客户端或者服务端出现问题的时候能正常使用是非常重要的。HDFS的容错能力大概可以分为两个方面：文件系统的容错性以及Hadoop本身的容错能力。文件系统的容错性通过这么几个手段：

（1）在目录节点和数据节点之间维持心跳检测。
（2）检测文件块的完整性。
（3）集群的负载均衡。
（4）维护多个FsImage和Editlog的拷贝。
（5）文件的删除。

从HDFS看分布式文件系统的设计需求

◘安全性问题

HDFS的安全性是比较弱的，只有简单的与unix文件系统类似的文件许可控制，未来版本会实现类似NFS的Kerberos验证系统。

本章小结

本章介绍Hadoop分布式文件系统HDFS，介绍了块、目录节点、数据节点等相关概念，并介绍了HDFS体系结构、命名空间、存储原理、通讯协议、数据错误和异常、尚未实现的功能总结，最后讲述了从HDFS看分布式文件系统的设计需求。

HDFS作为通用的分布式文件系统并不适合，它在并发控制、缓存一致性以及小文件读写的效率上是比较弱的。但是它有自己明确的设计目标，那就是支持大的数据文件（兆至T级），并且这些文件以顺序读为主，以文件读的高吞吐量为目标，并且与MapReduce框架紧密结合。

主讲教师和助教

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

助教：赖明星

单位：厦门大学计算机科学系数据库实验室2011级硕士研究生（导师：林子雨）
E-mail: mingxinglai@gmail.com
个人主页：http://mingxinglai.com

欢迎访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423
本讲义PPT存在配套教材《大数据技术基础》，请到上面网站下载。

Department of Computer Science, Xiamen University, Sep, 2013