大家好，在本节中，我们将向大家介绍一下HDFS体系结构中的相关概念。

(PPT2)

与单机的文件系统不同，HDFS文件系统不是将这些数据放在一块磁盘上，由上层操作系统来管理。而是存放在一个计算机集群上，由集群中的节点，各尽其责，通力合作，提供整个文件系统的服务。其中重要的节点包括：名称节点 (NameNode），数据节点（DataNode)及第二节点服务器（Secondary NameNode）。文件的目录结构独立存储在一个名称节点上，而具体文件数据，拆分成若干块，冗余的存放在不同的数据节点上。

这是HDFS的体系结构图，在学习该体系结构之前，让我们先了解数据块、名称节点、数据节点，第二名称节点相关概念。

（PPT3）

和传统的文件系统一样，为了提高磁盘的读写效率，HDFS也采用了数据块为单位来读写文件。HDFS中数据分块的另一个重要原因是：当一个文件大于集群中任意一个节点的时候，HDFS将文件分成数据块，不同的数据块分发到不同的节点中，这样一个文件的大小不会受到单个节点的容量限制，可以远远大于网络中任意节点的存储容量。

HDFS1.0的数据块的大小是64M。那么为什么如此设置呢？

（1）减少寻道时间。对于HDFS来讲，寻道是一个逻辑的概念，因为真正的寻道发生在磁盘，这里的寻道时间指的就是定位到块的时间。HDFS是存储大数据的，如果块设计的很小，一个文件就会由很多块组成，而HDFS上文件读写的最小单位是块，这样，寻找块的时间就会大大增加，降低读写效率。

（2）减少任务数。分布式计算是以一个块为单位处理的，如果块很小的话，mapreduce任务数就会非常多，任务之间的切换开销变大，效率降低，同样，如果块很大的话，一个任务中就会有很多很多的数据，这样任务就会很慢。

（3）适合数据备份。如果数据块很小，一个文件要分成很多块，而每个文件都有副本，当文件删除或者拷贝时，就会导致大量块移动，寻道开销和网络开销都会很大。

（PPT4）

从HDFS体系结构图看出，名称节点只有一个，在HDFS中身兼两职，一是服务客户端，二是管理各个数据节点（DataNode）。

对于客户端而言，名称节点（NameNode）上放着所有的文件目录信息，各个文件的分块信息，数据块的位置信息，要找一个文件，必须问问它，由此而得此名。

对于数据节点（DataNode）而言，它做为数据节点服务器的领导同志存在，管理各个数据节点服务器，收集它们的信息，了解所有数据节点的生存现状，然后给它们分配任务，指挥它们齐心协力为系统服务。

（PPT5）

在HDFS中，名称节点（NameNode）负责管理三类数据：文件系统的目录结构，文件的分块信息，数据块的位置信息（就数据块放置在哪些数据服务器上。

在HDFS的架构中，只有文件的目录结构和分块信息才会被持久化到本地磁盘上，这些信息以两种形式将文件永久保存在本地磁盘上：FsImage和EditLog。

FsImage：保存了最新的元数据检查点（文件系统的目录结构，文件的分块信息），长时间不更新。

EditLog：保存了HDFS中自最新的元数据检查点后的针对文件的创建、删除、重命名等操作，但是随着时间推移，Editlogs内存储的数据越来越多，导致运行速度越来越慢。

而数据块的位置信息，也就是数据块与数据节点的映射关系，仅仅存活在内存数据结构BlockMap中。名称节点需要周期性获取数据节点的心跳报告和块报告来更新BlockMap。

那么名称节点是怎样维护这些数据结构的呢？

名称节点在启动时，系统会将FsImage中的内容加载到内存中去，之后再执行EditLog中的操作，使得内存中的数据和实际同步，存在内存中来支持客户端的读。一旦在内存中成功建立文件系统元数据的映射，则创建一个新的FsImage文件和一个空的EditLog文件。名称节点启动之后，HDFS中的更新操作会重新写到EditLog文件中，因为FsImage文件一般都很大（GB级别的很常见），如果所有的更新操作都往FsImage文件中添加，这样会导致系统运行的十分缓慢，但是，如果往EditLog文件里面写就不会这样，因为EditLog 要小很多。

（PPT6）

在HDFS体系机构中，datanode是HDFS集群从节点，有很多个，主要负责存储文件的各个数据块（Block），而每一个block都可

以在多个datanode上存储多个副本（副本数量也可以通过参数dfs.replication设置，默认是3），这些数据块都被存储到数据节点的本地linux文件系统中。

数据节点还负责和客户端交互进行数据块的读写操作。

Datanode会定期向Namenode发送心跳信息，汇报自身所保存的block信息，namenode将这些信息汇总用于更新内存中的BlockMap文件，当发现某个文件的某个数据块的副本数量小于设置值是，名称节点就会指挥数据节点进行副本复制。

（PPT7）

在HDFS体系结构中，为什么设置第二名称节点呢？

第一个原因是EditLogs存放在内存中，随着时间更新操作越来越多，EditLogs越来越大，导致运行速度越来越慢。EditLogs过大时，名称节点合并fsimage和EditLogs，清空EditLogs。合并的时候名称节点就要重启，名称节点在启动过程中处于“安全模式”，对外只能提供读操作，不能写操作，因此影响了用户的使用。

第二个原因是第三个原因是名称节点只有一个，但它崩溃的时候，整个系统都不可以。第二名称节点作为名称节点的“检查点”，可以通过逐条执行EditLogs中的操作，将fsimage恢复到名称节点故障点时刻。