3.3HDFS的体系结构

（PPT1）

大家好，在本节中，我们将向同学们介绍HDFS的体系结构。

（PPT2）

从HDFS的体系结构图看，HDFS服务器集群主要由4类角色组成：名称节点（NameNode）、数据节点（DataNode）、客户端（Client）和第二名称节点（SecondaryNameNode）。他们各尽其责，通力合作，提供整个文件系统的服务。

（PPT3）

其中设计的最核心采用了主从结构（Master/Slave）：

名称节点（NameNode）是 Master，它是一个管理者。管理 HDFS 的命名空间，管理数据块（Block）映射信息，配置副本策略，处理客户端读写请求。

数据节点（DataNode）是Slave，名称节点（NameNode）下达命令，数据节点（DataNode）执行实际的操作。它负责存储实际的数据块，执行客户端（Client）发出的数据块读/写操作。同时定期向名称几点（NameNode）发送“心跳”信息，报告自己的状态。

而客户端要读数据的时候，首先将文件名发送给名称节点，名称节点根据文件名找到对应的块，再根据每个块信息找到实际存储各个数据块的数据节点的位置，并把数据节点位置发送给客户端，最后客户端直接访问这些数据节点获取数据。例如，客户端要读foo文件，首先把foo文件发送到名称节点上，名称节点获取了分块信息1,2,4，通过分块信息查找BlockMap，从而找到这些块分别放在哪些数据节点上，最后把这些数据块和数据节点信息发送给客户端。

客户端要写数据和读数据一样，首先访问名称节点，名称节点就会告诉客户端数据文件要分成多少块，每个块写到哪个数据节点上，然后客户端按照这些信息将数据分块写到制定的数据节点上。例如，客户端要存储bar文件，名称节点创建目录并根据文件的大小将bar文件分成2块，同时访问数据节点，看看哪些节点可以冗余保存这些块，最后将分块信息和数据节点信息发送给客户端。

当获取了数据块和数据块所在数据节点信息后，客户端就可以脱离名称节点，直接和数据节点交互，从而读写数据。

（PPT4）

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。客户端的主要功能有以下5个：

1、文件切分。客户端将文件上传 HDFS 的时候要根据名称节点的指示将文件切分成 一个一个的Block，然后进行到数据节点上存储。

2、与 NameNode 交互，获取文件的位置信息。

3、与 DataNode 交互，读取或者写入数据块。

4、可以通过HDFS提供一些命令来管理 HDFS，比如启动或者关闭HDFS，比如文件的上传、下载、复制、查看、格式化名称节点等。

5、可以通过HDFS提供的Java API编程访问HDFS。

（PPT5）

在HDFS1.0中，第二名称节点是名称节点的冷备份，当名称节点宕机时，第二名称节点不能马上替代名称节点工作，所以整个系统要暂停服务。此时第二名称节点将fsimage和EditLog的合并，然后将合并后的文件返回给NameNode，从而辅助恢复名称节点。

第二名称节点另一个重要的功能是辅助名称节点，分担其工作量。在名称节点正常运行期间，HDFS不断发生更新操作，这些操作都要写入到EditLog中，因此EditLog文件也会逐渐变大，这会影响名称节点的性能。因此第二名称节点定期合并 FsImage和EditLog到新的FsImage中，并推送给名称节点。在合并期间，名称节点用一个EditLog.new记录更新操作，当接收到第二名称节点推送的新的FsImage时，会用新的FsImage替换旧的FsImage文件，同时用EditLog.new文件替代EditLog文件，从而减少了EditLog

的大小。

（PPT6）

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。从图中可知，HDFS定义了5类通信协议，而5类通信协议都是建立在TCP/IP协议之上的，以此规范通信两端的约定。

1、ClientProtocol协议是客户端进程与Namenode进程之间进行通信所使用的协议，该协议接口定义了80多个方法，都是由客户端发起的，由Namenode响应的操作，例如：客户端要获取数据块信息、操纵HDFS的目录命名空间、打开与关闭文件流等

2、ClientDatanodeProtocol协议是客户端进程与Datanode进程之间进行通信所使用的协议；该协议接口定义数据块恢复的方法。

3、DatanodeProtocol协议是当Datanode进程需要与NameNode进程进行通信是需要基于此协议，例如发送心跳报告和块状态报告；

4、InterDatanodeProtocol协议是Datanode进程之间进行通信的协议，例如客户端进程启动复制数据块，此时可能需要在Datanode结点之间进行块副本的流水线复制操作。

5、NamenodeProtocol协议是SecondaryNameNode进程与Namenode进程进行通信的协议，例如对FsImage文件执行特定的操作。

（PPT7）

在HDFS1.0只设置唯一一个名称节点，这样做虽然大大简化了系统设计，但也带来了一些明显的局限性，具体如下：

1、命名空间的限制：名称节点的元数据是保存在内存中的，因此，名称节点能够容纳的元数据（文件、块）的个数会受到内存空间大小的限制。

2、性能的瓶颈：整个分布式文件系统的吞吐量，受限于单个名称节点的吞吐量。

3、隔离问题：由于集群中只有一个名称节点，只有一个命名空间，因此，无法对不同应用程序进行隔离。

4、集群的可用性：一旦这个唯一的名称节点发生故障，会导致整个集群变得不可用。