对比块存储、文件存储、对象存储的优缺点

块存储系统：

块存储，英文 block storage ，在一些地方也成为 raw block storage（裸块存储）。块存储所用的设备叫块设备（block device），常见的块设备有磁盘、闪存等。设备的存储都是按照一块块区域进行划分的，比如磁盘是按照盘面、扇区等将数据进行分开存储。

优点

1、通过RAID与LVM等手段，对数据提供了保护（RAID可实现磁盘的备份和校验，LVM可以做快照）；

2、RAID将多块廉价的硬盘组合起来，构建大容量的逻辑盘对外提供服务，性价比高；

3、写数据时，由于是多块磁盘组合成的逻辑盘，可以并行写入，提升了读写效率；

4、很多时候块存储采用SAN架构组网，传输速率以及封装协议的原因，使得传输速度与读写速率得到提升。

缺点

1、采用SAN架构组网时，需要额外为主机购买光纤通道卡，还要买光纤交换机，造价成本高；

2、不利于不同操作系统主机间的数据共享，因为操作系统使用不同的文件系统，格式化完成后，不同文件系统间的数据是无法共享的。

文件存储系统：

文件存储带有文件系统，主要是以文件的形式存放数据，能将所有的目录、文件形成一个有层次的树形结构来管理，通过“树”不断伸展的枝丫就能找到你需要的文件。存储协议主要是NFS、CIFS等，以统一命名空间的形式共享一个存储空间，能够支持成百上千的用户进行访问并上传下载文件，共享非常方便。

优点

1、造价较低：只需要普通机器和普通网络即可满足需求，不需要专用的SAN网络；

2、**方便文件共享。**

缺点

读写速率低，传输速率慢：以太网，上传下载速度较慢，另外读写操作都分布到单台服务器，与磁阵的并行写相比性能差距较大。

对象存储系统：

对象包含数据和元数据，每个对象都有一个唯一的“身份码”（对象ID）和“接入码”（Key），只有当“码”经过认证后，才能通过基于http协议的RESTful接口进行访问。不同于块存储和文件存储，对象是存在“桶”里的，桶就像万能的“百宝袋”，支持文件、照片、视频等不同类型的对象，而且再多的数据都能装得下。

优点

1、结合了块存储与文件存储的优点。

缺点

1、数据库等追求高性能的应用更适合采用块存储。

2、对象存储的成本比普通的文件存储还是较高。

阅读论文并回答问题

Q1：客户端读取HDFS系统中文件指定偏移量处的数据时，工作流程是什么

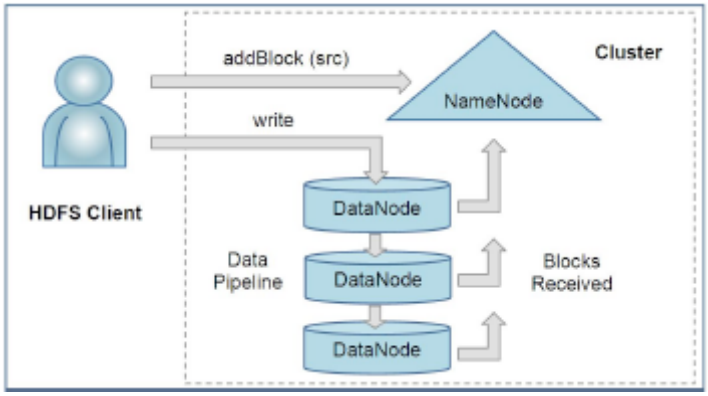
客户端首先向NameNode请求获得承载该文件块副本的DataNode列表。然后，它直接联系一个DataNode，请求传输所需的块。

当客户端打开一个文件进行读取时，它从NameNode获取区块的列表和每个区块副本的位置。每个区块的位置是按照它们与读者（客户端）的距离排序的。当读取一个区块的内容时，客户端首先尝试最近的副本。如果读取尝试失败，客户端将依次尝试下一个副本。如果目标数据节点不可用，该节点不再托管该块的副本，或者在测试校验和时发现副本损坏，则读取可能失败。

Q2：客户端向HDFS系统中指定文件追加写入数据的工作流程是什么

客户端写入时，它首先要求NameNode选择DataNodes来托管文件的第一个块的副本。客户端构建了一个从节点到节点的管道，并发送数据。当第一个区块被填满时，客户端要求选择新的DataNodes来承载下一个区块的副本。

打开一个文件进行写入的HDFS客户端被授予该文件的租约；其他客户端不能写入该文件。写入的客户端通过向NameNode发送心跳来定期更新租约。当文件被关闭时，租约被撤销。租约持续时间由软限制和硬限制约束。在软限制过期之前，写作者确定对文件的独占访问。如果软限制过期，而客户未能关闭文件或续租，另一个客户可以抢占租约。如果在硬限制到期后（一小时），客户端未能更新租约，HDFS会认为客户端已经退出，并会代表写作者自动关闭文件，并恢复租约。写作者的租约并不妨碍其他客户端阅读该文件；一个文件可能有许多并发的读者。



Q3：新增加一个数据块时，HDFS如何选择存储该数据块的物理节点?

当需要一个新的区块时，NameNode分配一个具有唯一区块ID的区块，并确定一个DataNode列表来承载该区块的副本。 数据节点形成一个管道，其顺序成一条管道，其顺序使从客户端到最后一个数据节点的总网络距离最小。字节作为数据包序列被推送到管道中。应用程序首先写入的字节在客户端进行缓冲。在一个数据包缓冲区被填满后（通常是64KB），数据被推送到管道。在收到前一个数据包的确认之前，下一个数据包可以被推送到管道中。未处理的数据包的数量受到客户端未处理数据包窗口大小的限制。

Q4: HDFS采用了哪些措施应对数据块损坏或丢失问题?

损坏：HDFS为HDFS文件的每个数据块生成并存储校验和。校验和由HDFS客户端在读取时进行验证，以帮助检测由客户端、数据节点或网络造成的任何损坏。当客户端创建一个HDFS文件时，它为每个块计算校验和序列，并将其与数据一起发送给DataNode。DataNode将校验和存储在元数据文件中，与块的数据文件分开。当HDFS读取一个文件时，每个区块的数据和校验和被送到客户端。客户端计算收到的数据的校验和，并验证新计算的校验和是否与它重新收到的校验和相符。如果不匹配，客户端会通知NameNode损坏的副本，然后从另一个DataNode获取该块的不同副本。同时，NameNode开始复制该区块的一个良好副本。只有当良好的副本数量达到该区块的副本因子时，才会安排删除损坏的副本。

丢失：将数据复制三次，防止由于不相关的节点故障而导致数据丢失。

对于大型集群，每天都会有一两个节点丢失。同一个集群将在大约两分钟内重建托管在故障节点上的54000个块复制。（重新复制的速度很快，因为这是一个并行问题）。（在两分钟内有数个节点发生故障，导致某个区块的所有复制都丢失的概率非常小）

Q5：HDFS采用什么措施应对主节点失效问题

HDFS引入了备份节点BackupNode，BackupNode能够创建周期性的检查点，同时维护着一个与NameNode的状态同步的一个图像。备份节点接受来自激活的NameNode的命名空间事务日志流，将其保存到自己的存储目录中，并将这些事务应用于内存中自己的命名空间图像。NameNode将BackupNode视为日志存储，就像它对待其存储目录中的日志文件一样。如果NameNode发生故障，BackupNode在内存中的映像和磁盘上的检查点是最新命名空间状态的记录。

Q6: NameNode 维护的“数据块 物理节点对应表”需不需要在硬盘中备份？为什么?

不需要。HDFS将整个命名空间存放在RAM中。在启动期间，NameNode通过读取命名空间和重放日志来恢复命名空间。当“数据块—物理节点对应表”失效时可通过向NameNode请求得到最新的文件块位置信息。