三类分布式存储系统的区别

（1）块存储系统

块存储是将裸磁盘空间整个映射给主机使用的，比如磁阵中有 3 块 1T 硬盘，可以选择直接将裸设备给操作系统使用（此时识别出 3 个 1T 的硬盘），也可以划分经过 RAID、逻辑卷等方式划分出多个逻辑的磁盘供系统使用（比如划分为 6 个 500G 的磁盘），主机层面操作系统识别出硬盘，但是操作系统无法区分这些映射上来的磁盘到底是真正的物理磁盘还是二次划分的逻辑磁盘，操作系统接着对磁盘进行分区、格式化，与我们服务器内置的硬盘没有什么差异。

块存储不仅仅是直接使用物理设备，间接使用物理设备的也叫块设备，比如虚拟机创建虚拟磁盘。VMware、VirtualBox 都可以创建虚拟磁盘，虚拟机创建的磁盘格式包括 raw、qcow2 等，这与主机使用的裸设备不一样，且有不同的应用场景。

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。

（2）对象存储系统

之所以出现对象存储，是为了克服块存储与文件存储的缺点，发扬他俩各自的优点。简单地说，块存储读写块，不利于共享，文件存储读写慢，利于共享。

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。

（3）文件存储系统

为了克服块存储无法共享的问题，所以就有了文件存储。

文件存储也有软硬一体化的设备，用一台普通服务器/笔记本，只要安装上合适的操作系统与软件，就可以对外提供 FTP 与 NFS 服务。

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

1、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。

客户端读取 HDFS 系统中指定文件制定偏移量处的数据时,工作流程是什么?

HDFS 客户端首先向 NameNode 询问承载该文件块副本的 DataNode 列表。然后，它直接与 DataNode 联系，并请求传输所需的块。

客户端打开要读取的文件时，它将从 NameNode 获取块列表和每个块副本的位置。每个块的位置按它们与阅读器的距离排序。读取块的内容时，客户端首先尝试最接近的副本。如果读取尝试失败，则客户端将依次尝试下一个副本。如果目标 DataNode 不可用，该节点不再托管该块的副本或在测试校验和时发现该副本已损坏，则读取可能会失败。

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？

（1）打开文件进行写入的 HDFS 客户端将获得文件的租约；没有其他客户端可以写入文件。

（2）应用程序在客户端写入第一个缓冲区的字节。填充数据包缓冲区（通常为 64 KB）后，数据将被推送到管道。在接收到先前数据包的确认之前，可以将下一个数据包推送到管道。未完成数据包的数量受客户端未完成数据包窗口大小的限制。

（3）在将数据写入 HDFS 文件之后，HDFS 不会提供任何保证，直到关闭文件后新读取器才能看到该数据。如果用户应用程序需要可见性保证，则可以显式调用 hflush 操作。然后将当前数据包立即推送到管道，并且 hflush 操作将等待，直到管道中的所有 DataNode 都知道成功传输了数据包。这样，在进行 hflush 操作之前写入的所有数据都肯定对读者可见。

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？

创建新块时，HDFS 将第一个副本放置在写入程序所在的节点上，将第二个和第三个副本放置在不同机架中的两个不同节点上，其余放置在随机节点上，但限制不超过当副本数少于机架数的两倍时，在一个节点上放置一个副本，在同一

机架中放置不超过两个副本。将第二个和第三个副本放置在不同机架上的选择可以更好地在整个群集中分配单个文件的块副本。如果前两个副本放在同一机架上，则对于任何文件，选择了所有目标节点之后，按照它们与第一个副本的接近程度的顺序将节点组织为管道。数据按此顺序推送到节点。

HDFS 采用了哪些措施应对数据块损坏或丢失问题？

每个 DataNode 运行一个块扫描器，该扫描器定期扫描其块副本并验证存储的校验和是否与块数据匹配。每当读取客户端或块扫描器检测到损坏的块时，它都会通知 NameNode。NameNode 将副本标记为已损坏，但不会立即计划删除副本。相反，它开始复制该块的良好副本。仅当良好的副本数达到块的复制因子时，才计划删除损坏的副本。此政策旨在尽可能长时间地保留数据。因此，即使块的所有副本都已损坏，该策略也允许用户从损坏的副本中检索其数据。

HDFS 采用了什么措施应对主节点失效问题？

引入 BackupNode。BackupNode 能够创建定期的检查点，但除此之外，它还维护文件系统名称空间的内存中最新映像，该映像始终与 NameNode 的状态同步。如果 NameNode 失败，则内存中 BackupNode 的映像和磁盘上的检查点是最新名称空间状态的记录。

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。
因为文件块位置信息只存储在内存中，是在 DataNode 加入集群的时候，NameNode 询问 DataNode 得到的，并且间断的更新。所以当“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。