块存储系统、对象存储系统、文件存储系统的区别？

块存储系统：块是指以扇区为基础，一个或多个连续的扇区组成一个块，也称物理块，因此使用块存储 IO 效率高、实时性强。块存储系统可以对物磁盘空间进行分割，存储数据时以块为单位进行存储，将相邻的两块划分到不同的物理盘上，可以起到保护数据的作用，避免因一块磁盘损坏而导致全部数据丢失。块存储系统将多块廉价的硬盘组合起来，成为一个大容量的逻辑盘对外提供服务，提高容量。多个块可以并行读写，提高数据吞吐效率。

文件存储系统：单个文件可能由于一个或多个逻辑块组成，且逻辑块之间是不连续分布，逻辑块大于或等于物理块整数倍。读取某些文件时，首先会查找每个文件逻辑块，其次物理块，由于逻辑块是分散在物理块上，而物理块也是分散在不同扇区上。需要一层一层查找，最后才完成整个文件读取，比较费时，效率不高，实时性不强。但是不必架设专用网络，使用普通以太网即可，因此文件存储系统具有造价低廉、方便共享的优点。

对象存储系统：对象是系统中数据存储的基本单位，一个对象实际上就是文件的数据和一组属性信息（Meta Data）的组合，这些属性信息可以定义基于文件的 RAID 参数、数据分布和服务质量等。在存储设备中，所有对象都有一个对象标识，通过对象标识 OSD 命令访问该对象。对象存储同兼具 SAN 高速直接访问磁盘特点及 NAS 的分布式共享特点，核心思想是将数据通路（数据读或写）和控制通路（元数据）分离，并且基于对象存储设备构建存储系统，每个对象存储设备具有一定的智能，能够自动管理其上的数据分布。对象存储一般不支持追加写和更新，面向的是一次写入，多次读取的需求场景

阅读论文《The Hadoop Distributed File System》并回答一些问题：

1、端读取 HDFS 系统中指定文件指定偏移量处的数据时，工作流程是什么？

HDFS 客户端首先向名称节点（NameNode）询问保存该文件块副本的数据节点（DataNode）列表。然后，客户端直接与 DataNode 联系，并请求传输所需的块。具体来说，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。

2、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）**申请租约** 申请写入数据的 HDFS 客户端将获得文件的租约，此时其他客户端不可以写入该文件。

（2）**写入数据** 应用程序在客户端将数据写入缓冲区（通常为 64 KB）后，缓冲区填满之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，未完成确认的数据包的数量受客户端窗口大小的限制。

（3）**刷新数据** 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，则可以显式调用刷新（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才对读取者可见。

3、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。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。

4、HDFS 采用了哪些措施应对数据块损坏或丢失问题？

每个 DataNode 都运行一个块扫描器，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。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。

5、HDFS 采用了什么措施应对主节点失效问题？

采用主从备份，备份节点（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。这样即使主节点崩溃失效，系统也可以切换至从节点继续工作。

6、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。