**一、说明三类分布式存储系统的区别：(1)块存储系统； (2)对象存储系统；(3)文件存储系统。**

**块存储：这种接口通常以QEMU Driver或者Kernel Module的方式存在，这种接口需要实现Linux的Block Device的接口或者QEMU提供的Block Driver接口，如Sheepdog，AWS的EBS，青云的云硬盘和阿里云的盘古系统，还有Ceph的RBD。主要可以应付热点问题。**

**对象存储：其接口就是简单的GET,PUT,DEL和其他拓展，一个对象我们可以看成一个文件，只能全写全读，通常以大文件为主，要求足够的IO带宽**

**文件存储：通常意义是支持POSIX接口，它跟传统的文件系统如Ext4是一个类型的，但区别在于分布式存储提供了并行化的能力，，但是有时候又会把GFS，HDFS这种非POSIX接口的类文件存储接口归入此类，比块存储具有更丰富的接口，需要考虑目录、文件属性等支持**

**阅读《The**[Hadoop](https://so.csdn.net/so/search?q=Hadoop&spm=1001.2101.3001.7020)**Distributed File System》回答问题**

#### 客户端读取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？

**当客户端打开文件并读取，它会从NameNode获取块列表和每个块副本的位置。位置信息会根据离reader的距离的顺序排序。读取一个块的内容，客户端会尝试从最近的副本读。如果失败了，就读取序列中的下一个副本。**

#### 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，其他客户端都不能写入文件。正在写入的客户端发送给NameNode来定期更新租约。当文件被关闭，租约就被撤销。**
2. **字节作为分组序列被推入流水线。应用程序在客户端写入第一个缓冲区，填充完一个缓冲区后，数据被推送到流水线。在接收到先前数据包的确认信息之前，可以将下一个分组推送到流水线。**
3. **在数据被写入HDFS文件后，HDFS不保证数据对新reader可见，直到文件被关闭。**

#### 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？

**当一个新块被创建，HDFS把第一份副本存放到writer所在的节点，第二和第三份存放到不同机架的不同节点，当副本数小于两倍机架数时，每个节点上的副本不多于1份，每个机架上的副本不多于2份。把第二第三份副本放到不同机架能够在集群中更好地分发单个文件的块副本。如果任意文件的头两份副本放在同一个机架，2/3的块副本会在同一个机架上。**

#### HDFS采用了哪些措施应对数据块损坏或丢失问题？

**当读客户端或block scanner检测到坏块时，会通知NameNode。NameNode将块标记为损坏，但是不会马上安排删除副本，而是复制出一个好的副本。仅在好副本数达到复制因子，损坏的副本才会被安排删除。**

1. **HDFS采用了什么措施应对主节点失效问题？**

**如果NameNode发生了故障，BackupNode内存中的image和磁盘上的checkpoint就是最新的命名空间状态。**

#### ⑥NameNode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HDFS将整个命名空间存放在RAM中。  
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。**