1. 三类分布式存储系统的区别:
2. 块存储系统
3. 对象存储系统
4. 文件存储系统

存储接口区别：

1.对象存储：也就是通常意义的键值存储，其接口就是简单的GET,PUT,DEL和其他扩展，如七牛、又拍，Swift，S3

2.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(RBD是Ceph面向块存储的接口)

3.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，如Ceph的CephFS(CephFS是Ceph面向文件存储的接口)，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。

IO特点区别：

1.对象存储(键值数据库)：接口简单，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。

2.块存储(硬盘)：它的IO特点与传统的硬盘是一致的，一个硬盘应该是能面向通用需求的，即能应付大文件读写，也能处理好小文件读写。但是硬盘的特点是容量大，热点明显。因此块存储主要可以应付热点问题。另外，块存储要求的延迟是最低的。

3.文件存储(文件系统)：支持文件存储的接口的系统设计跟传统本地文件系统如Ext4这种的特点和难点是一致的，它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，实现一个支持并行化的文件存储应该是最困难的。但像HDFS、GFS这种自己定义标准的系统，可以通过根据实现来定义接口，会容易一点。

1. The Hadoop Distributed File System
2. 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?

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。

1. 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（lease），其他客户端都不能写入文件。正在写入的客户端通过发送心跳给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（通常64KB）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。未完成分组的数量受客户端未完成分组的窗口大小的限制。

（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。如果用户应用需要这种可见性保证，可以显式调用 hflush 操作。这样当前的分组会立即推送到流水线，hflush操作会等待所有的DataNode确认已经成功传输了分组。所有之前写入的数据就会理所应当地对reader可见。

1. 新增一个数据块时，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。

1. 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。

1. 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发生了故障，BackupNode内存中的映像和磁盘上的checkpoint就是最新的命名空间状态。

1. 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。