# 三类分布式存储系统的区别

## 存储接口的区别：

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。

因此，这三种接口分别以非分布式情况下的键值数据库、硬盘和文件系统的IO特点来对应即可。至于冷热、快慢、大小文件而言更接近于业务。但是因为存储系统是通用化实现，通常来说，需要尽量满足各种需求，而接口定义已经一定意义上就砍去了一些需求，如对象存储会以冷存储更多，大文件为主。

# 阅读论文《The Hadoop Distributed File System》回答问题

#### **客户端读取[HDFS](https://so.csdn.net/so/search?q=HDFS&spm=1001.2101.3001.7020" \t "_blank)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。当读取一个块的内容，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。读取操作会因为目标DataNode不可用而失败、节点不再托管块的副本、测试校验和后发现块副本损坏而失败。

## 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（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。

## 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**？

1. 每个DataNode上都允许了一个block scanner，定期地扫描块副本并验证相关的校验和
2. 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。这个策略旨在尽可能久地保存数据。所以如果所有的副本都损坏了，该策略允许用户从损坏的副本中读取数据。

## HDFS采用了什么措施应对主节点失效问题？

**BackupNode**

1. 和CHeckpointNode类似，BackupNode也有能力创建定期检查点，但除此之外还在内存中维护文件系统命名空间最新的image，这总是和NameNode的状态一起同步
2. 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。

## 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。

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。