# 第五次分布式作业

## 三类存储系统的区别：

1、块存储

可以看作为是裸盘，最明显的特点是不可以被操作系统直接浏览。可以凭借划分逻辑卷、做RAID、LVM等方式将它格式化，可以格式化为你所指定的文件系统（Ext3，Ext4，NTFS，FAT32等），随后才可以被操作系统浏览。比较常见的DAS、FC-SAN、IP-SAN基本都是块存储。它的优势是：读写速度快（带宽和IOPS）；可是缺点也一样明显：太过于底层，不利扩展，不可以被共享。

2、文件存储

文件存储可以分成本地文件存储和网络文件存储。文件存储最明显的特点是支持POSIX的文件浏览接口：open、read、write、seek、close等；优势是：有利于扩展&共享；缺点是：读写速度比较慢。

3、对象存储

有种表述是对象存储一定是分布式存储，其实不是。这是因为分布式文件存储可能是分布式存储，而不一定是对象存储。比较常见的对象存储有摩杜云的MOS，而实际对象存储是什么，可以去看摩杜云的帮助文档。

三者的数据“使用者”来进行区别：

1.块存储的“使用者”是可以读写块设备的软件系统，比如EXT3、EXT4等传统文件系统，或是数据库。

2.文件存储的“使用者”是我们操作人员，真真正正的自然人。

3.对象存储的“使用者”则是其它的计算机软件，比如OpenStack里调用镜像文件的glance模块。

4.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

5. 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)

6.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。

## 论文有关问题：

1. 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?
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?
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?
4. HDFS采用了哪些措施应对数据块损坏或丢失问题?
5. HDFS采用了什么措施应对主节点失效问题?
6. HDFS采用了哪些措施应对数据块损坏或丢失问题?
7. 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?

①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。

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）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。

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，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。

④HDFS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。

⑤HDFS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。

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。