第五次作业

学生：高国豪 学号：19030100397

老师：李龙海

回答下列问题：

1. 说明三类分布式存储系统的区别：

(1)块存储系统；(2)文件存储系统；(3)对象存储系统。

（1）块存储系统

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）。

块级是指以扇区为基础，一个或我连续的扇区组成一个块，也叫物理块。它是在文件系统与块设备（例如：磁盘驱动器）之间。

块存储会将数据拆分成块，并单独存储各个块。每个数据块都有一个唯一标识符，所以存储系统能将较小的数据存放在最方便的位置。块存储是底层存储，直接写入或读取硬盘扇区（块）。

（2）文件存储系统

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。

文件级是指文件系统，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文件存储会以文件和文件夹的层次结构来整理和呈现数据。

文件存储的用户是自然人，最容易理解。所有用于同一用途的数据，按照不同应用程序要求的结构方式组成不同类型的文件，然后我们给每一个文件起一个方便理解记忆的名字。而当文件很多的时候，我们按照某种划分方式给这些文件分组，每一组文件放在同一个目录（或者叫文件夹）里面，当然我们也需要给这些目录起一个容易理解和记忆的名字。而且目录下面除了文件还可以有下一级目录（称之为子目录或者子文件夹），所有的文件、目录形成一个树状结构。

（3）对象存储

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。

对象存储（Object-based Storage）是一种新的网络存储架构。

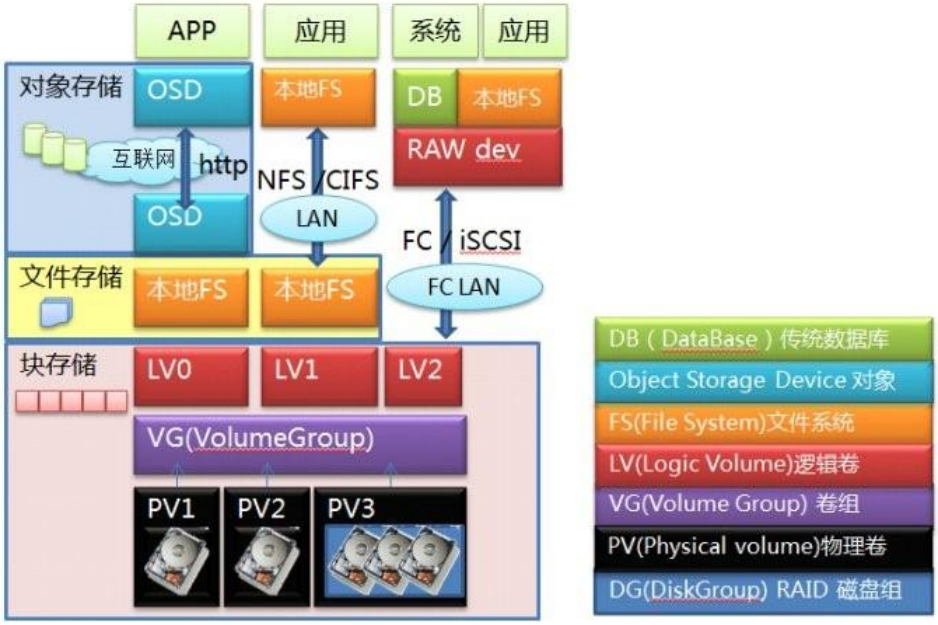
对象存储，也称为基于对象的存储，是一种扁平结构，其中的文件被拆分成多个部分并散布在多个硬件间。在对象存储中，数据会被分解为称为“对象”的离散单元，并保存在单个存储库中，而不是作为文件夹中的文件或服务器上的块来保存。

对象存储卷会作为模块化单元来工作：每个卷都是一个自包含式存储库，均含有数据、允许在分布式系统上找到对象的唯一标识符以及描述数据的元数据。元数据很重要，其包括年龄、隐私/安全信息和访问突发事件等详细信息。对象存储元数据也可以非常详细，并且能够存储与视频拍摄地点、所用相机和各个帧中特写的演员有关的信息。为了检索数据，存储操作系统会使用元数据和标识符，这样可以更好地分配负载，并允许管理员应用策略来执行更强大的搜索。

内容比较：

	块存储	文件存储	对象存储
存储设备	磁盘阵列，硬盘，虚拟硬盘	FTP、NFS 服务器，SamBa	内置大容量硬盘的分布式服务器
优点	读写速度最快	人可以直接使用，容易管理，价格便宜	读写速度和块存储相当，查询速度最快，扩容简单，程序容易管理，安全性较高。
缺点	但查询速度最慢，管理查询难	读写速度最慢，安全性较差	无法修改对象，必须一次性完整地写入对象

层次关系：



2.阅读论文《The Hadoop Distributed File System》并回答下面问题：

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？

HDFS 客户端第一步向 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。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）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。如果用户应用程序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管道，并且 hflush 操作会等待，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 reader 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。

1. 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。
2. 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。