第五次作业

19030100295 尚丹彤

三类分布式存储系统的区别

1. 块存储系统

   块存储提供的是不带文件系统裸磁盘，使用之前需先进行初始化。块是系统的基本存储单位，如机械硬盘的一个扇区。块存储只关心数据的进来和出去，不关心数据之间的关系和用途。由于块存储只负责数据读取和写入，因此具有有高带宽、低延迟的优势，但是扩展能力有限，适用于对响应时间要求高的系统。

2. 文件存储系统

   文件是系统的基本存储单位。文件存储一般体现形式是目录和文件，数据以文件的方式存储和访问，按照目录结构进行组织。其索引和数据是存放在一起的。文件存储的存储端带有文件系统，这些文件存储设备除了磁盘外还带有文件系统，用户直接通过存储端的文件系统就能调用存储资源。相比于块存储，文件存储读写速度相对于块存储要慢一点。

3. 对象存储系统

   对象是系统的基本存储单位，对象可以是文件的某一块。其索引和数据的分开独立存储的。对象存储端的文件系统采用类似哈希表-键值的方式来提高读写速度。对象存储就可以非常简单的扩展到超大规模，因此非常适合数据量大、增速又很快的视频、图像等，例如百度网盘。

论文解读

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客户端先向NameNode询问承载该文件块副本的DataNode列表。各个数据块按照它们与阅读器的距离排序。当读取数据块的内容时，客户端首先会尝试距离最近的块副本。当读取失败时，客户端会依次尝试该块的下一个副本。如果目标DataNode不可用，该节点不再托管该块的副本或读取到的副本的 checksum 错误，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?

打开文件进行写入的HDFS客户端将会获得文件的租约，此时其他客户端不能写入文件。一个HDFS文件由若干数据块组成。当需要新的数据块时，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。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。

3. 当新增一个数据块时，HDFS 如何选择存储该数据块的物理节点?

当一个新的数据库被创建时，HDFS 将在写入节点上创建第一个副本，并在处于其他机架上的两个不同节点创建第二和第三个副本，其他副本随机将会被随机放置。并且满足: 至多有一个副本被放置在同一结点， 当副本数量小于机架数量的两倍时，至多有两个副本被放置在同一机架。

4. HDFS 采用了哪些措施应对数据块损坏或丢失问题?

HDFS 通过副本和 checksum 的方式应对数据快损坏或丢失问题。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。 这一措施旨在尽可能地保留数据，使得即使某数据块的所有副本都损坏了，用户仍可能从损坏的副本中修复数据。

5. HDFS 采用了什么措施应对主节点失效问题?

正常运行时期间，DataNode 会向 NameNode 发送心跳以确保 DataNode 正在工作并且其 主管的数据块副本是可用的。默认的心跳间隔为 3*s*。如果 NameNode 在十分钟内都没有收到来 自 DataNode 的心跳，NameNode 会认为该 DataNode 停止了服务并且该 DataNode 所主管的数据 块副本不可用。此时，NameNode 会计划在其他 DataNode 上创建这些数据块的新的副本

6. NameNode 维护的“数据块—物理结点对应表”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。