分布式第五次作业

一、自己查阅资料，说明三类分布式存储系统的区别：

（1）块存储系统；（2）对象存储系统；（3）文件存储系统。

（1）块存储系统

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。块存储通常会被配置为将数据与用户环境分离，并会将数据分布到可以更好地为其提供服务的多个环境中。然后，当用户请求数据时，底层存储软件会重新组装来自这些环境的数据块，并将它们呈现给用户。

由于块存储不依赖于单条数据路径（和文件存储一样），因此可以实现快速检索。每个块都独立存在，且可进行分区，因此可以通过不同的操作系统进行访问，这使得用户可以完全自由地配置数据。它是一种高效可靠的数据存储方式，且易于使用和管理。

（2）文件存储系统

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（通常用不同的后缀来指代不同的类型），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。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。当要访问文件时，根据目录和文件名就能访问到我们想要访问的文件。

但是，文件存储系统不仅是依赖于单条数据路径。而且基于文件的存储系统必须通过添置更多系统来进行横向扩展，而不是通过增添更多容量来进行纵向扩展。

（3）对象存储系统

对象存储一般体现形式是一个UUID，数据和元数据打包在一起作为一个整体对象存在一个超大池子里。对于对象访问，只需要报出它的UUID，就能立即找到它，但访问的时候对象是作为一个整体访问的。

对象存储将元数据独立出来了，控制节点叫元数据服务器（服务器+对象存储管理软件），里面主要负责存储对象的属性（主要是对象的数据被打散存放到了那几台分布式服务器中的信息）。而其他负责存储数据的分布式服务器叫做OSD，主要负责存储文件的数据部分。当用户访问对象，会先访问元数据服务器，元数据服务器只负责反馈对象存储在哪几个OSD，假设反馈文件A存储在B、C、D三台OSD，那么用户就会再次直接访问3台OSD服务器去读取数据。

	块存储系统	**文件存储系统**	**对象存储系统**
体现形式	卷或硬盘	目录或文件	UUID
数据访问单位	字节	文件	对象（元数据+数据）
一致性	最终一致性	强一致性	强一致性
典型设备	磁盘阵列、硬盘	FTP、NFS服务器	内置大容量硬盘的分布式服务器
优点	读写快	利于共享	读写快、利于共享

二、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?

1. 客户端向NameNode发送读请求（文件名，偏移量，长度）；
2. NameNode根据文件名、偏移量找到对应的DataNode，并块列表和每个块副本的位置发送个客户端；
3. 客户端根据就近原则选择某个数据节点，进行读取数据。

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. 客户端向NameNode发送写请求；
2. NameNode应答；
3. 客户端向NameNode发送写入第一个数据块的请求；
4. NameNode根据负载均衡策略选择3个DataNode，并将其对应的IP列表返回给客户端；
5. 客户端将这3个DataNode构成一个流水线，将第一个数据块的数据流写入流水线；
6. 第一个数据块写入成功之后再向NameNode获取下一个数据块对应的3个NameNode。

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?

当新增加一个数据块时，HDFS将第一个副本放在当前写入的DataNode，第二个和第三个副本放在不同机架的不同节点上。当副本数小于两倍机架数时，每个节点上的副本不多于一份，每个机架上的副本不多于两份。把第二份和第三份副本放在不同的机架上能够在集群中更好的分发单个文件的块副本。

④HDFS采用了哪些措施应对数据块损坏或丢失问题?

1. 数据块损坏

   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。

   当客户端或block scanner检测到损坏的数据块时，都会通知NameNode。NameNode将该副本标记为损坏，但不会立即计划删除该副本。相反，它开始复制块的一个良好副本。只有当良好的副本计数达到块的复制因子时，损坏的副本才计划被移除。该策略旨在尽可能长时间保存数据。因此即使一个块的所有副本损坏了，该策略也允许用户从损坏的副本中检索数据。

2. 数据块丢失

   客户端发送数据时是以数据块为单位按顺序发送的，若接受完毕，DataNode会向客户端发送ack确认码。

⑤HDFS采用了什么措施应对主节点失效问题?

与 checkpointnode 类似，backupnode 能够创建周期性的检查点，但是除此之外，它还在内存中维护文件系统名称空间的最新映像，该映像始终与 namenode 的状态同步。

如果 namenode 失败，则备份节点在内存中的映像和磁盘上的检查点是最新命名空间状态的记录。

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DataNode不断向NameNode发送存储的数据块信息，"数据块——物理节点对应表"是在不断更新的。所以不需要备份。