谷歌的三架马车

--19301161-喻鑫

1 《Google File System》

       GFS是一个分布式文件系统，用来存储大量的较大文件，它可以在廉价的硬件上实现存储文件，并做到容错性，并且针对多个客户同时访问提供比较有竞争力的性能。这篇文章主要讲的是Google的GFS文件系统，一个面向大规模数据密集型应用的、可伸缩的分布式文件系统。它与传统的分布式系统有着很多相同的设计目标，比如，性能、可伸缩性以及可用性，并且它还考虑了负载均衡问题和技术环境的影响。设计思路：组件失效是常态事件，由于其系统是由几百甚至几千台普通廉价设备组装而成的存储机器，因此在使用过程中难免会出现各种各样的问题。文件巨大，以TB为数量级的数据的存储。大多数文件采用文件尾部追加数据而不是覆盖原有数据。应用程序和文件系统API的协同设计提高了整个系统的灵活性。由于放松了GFS的一致性的要求，简化了GFS的设计，提高了系统的灵活性。  
中间部分主要讲的是GFS文件系统的设计阶段，涉及到它要达到的预期效果 、接口的设计、架构设计、节点设计、元数据的结构与存储等主要的设计过程。过程中还包含了操作日志的记录与查看以及GFS文件系统为了达到它支持的一致性而设计的一些机制规则，比如有一致性保障机制、程序实现。在系统交互方面，它的原则就是最小化所有操作和Master节点的交互。那他使用的规则就是租赁和变更顺序、数据流、原子记录追加和快照，旨在数据更好地交互与调用。再一个讲的比较多的就是Master节点的操作，它执行所有的名字空间操作，管理整个系统里所有Chunk副本，Chunk副本有三个用途，分别是Chunk创建、重新复制和重新负载均衡。在垃圾回收方面，GFS文件删除不会立刻回收可用的物理空间，它是采用惰性回收的策略。

2 《Google Bigtable》

BigTable是一个分布式存储系统，可以管理结构化数据。它支持水平扩展，并通过使用数千台连接的服务器支持PB级数据处理。稍后，我们将介绍它的一些组件及其数据模型。首先，我们可以看到数据模型。BigTable是一个稀疏、分布式和持久的多维排序映射，该映射包含行关键字、列关键字和时间戳。BigTable还提供API函数来修改集群、表和列族的元数据。它的组件基于几个Google基本组件。它使用谷歌的分布式系统GFS来存储日志文件和数据文件。它还引入了BigTable所依赖的高可用性和序列化分布式锁服务组件，称为chubby。BigTable主要包括三个组件：链接到客户端程序的库、主服务器和多个tablet服务器。根据任务的变化，它还可以动态地将平板服务器添加到集群中。此外，谷歌采用类似于B+树的三层结构来存储平板电脑的位置信息，位置信息不超过三层。例如，平板电脑分配意味着平板电脑只能分配给一台平板电脑服务器。文章还详细介绍了平板电脑服务及其内存空间回收机制

上面主要讨论Google BigTable的实现组件及其底层框架。本段主要介绍其优化工作和一些性能数据。首先是在优化方面所做的努力，（1）它使用一个本地组，即客户端程序可以将多个列族组合成一个本地组。将为平板电脑中的每个本地组生成单独的sstable。将不能一起访问的列族划分为不同的本地组可以提高读取操作的效率。（2） 使用压缩技术，客户机程序可以控制本地组的sstables是否需要压缩。当然，压缩可以在不影响读取速度的情况下减少存储空间。（3） 缓存用于提高读取操作的性能。平板电脑服务器使用辅助缓存策略。扫描缓存为主缓存，主要缓存平板服务器通过sstable接口获取的键值对；块缓存是辅助缓存，用于缓存从GFS读取的sstable块。例如，bloom过滤器用于实现提交日志恢复速度和不变性。其次，在性能评估方面，为了测试BigTable的性能和可扩展性，Google建立了一个包含n台平板服务器的BigTable集群，其数量是可变的。

3 《Google MapReduce》

MapReduce是一种用于处理大型数据集的编程模型。它通过map函数将基于行的输入转换为不同的键值对，然后通过reduce函数为同一个键聚合这些键值对。。MapReduce架构程序可以在大量普通机器上执行并行处理。它可以在普通计算机集群上运行，集群规模可以灵活调整。它仍然使用大量普通机器来形成集群以提高其性能。为了解决并行计算、数据分布和处理错误等问题，Google受到了LISP语言的启发，可以编写map和reduce的原语。

中间部分主要介绍MapReduce的原理及其使用环境。首先介绍了MapReduce的编程模型原理。它主要使用输入集生成输出集。与MapReduce库相对应，使用了两个函数来表示计算，即map和reduce。MapReduce模型可以以多种不同的方式实现。根据实际情况进行配置。在谷歌内部，它是由多台普通PC通过以太网连接而成的集群实现的。在执行操作时，map和reduce都可以将数据分成多个段，然后部署到多台机器上运行，从而达到并行计算的目的。数据结构使用主数据结构，主数据结构存储数据的大小和位置。MapReduce还具有良好的容错机制。例如，如果工人失败，如果map任务已完成，则将再次访问该任务；如果reduce任务已完成，则无需再次执行。然后是主机故障和数据故障的容错处理。在MapReduce操作期间，数据存储在GFS上。除了map和reduce两种处理功能外，MapReduce还具有许多扩展功能，如划分功能、序列保证功能和合并功能。它还支持各种不同格式的输入数据。