05 | 从RAID看垂直伸缩到水平伸缩的演化

2018-11-08 李智慧

从0开始学大数据 进入课程 >

讲述：李智慧
时长 10:07 大小 4.64M

经过前面的预习和上一期我们聊的，大数据技术主要是要解决大规模数据的计算处理问题，但是我们要想对数据进行计算，首先要解决的其实是大规模数据的存储问题。我这里有一个直观又现实的问题想问你：如果一个文件的大小超过了一张磁盘的大小，你该如何存储？

我的答案是，单机时代，主要的解决方案是 RAID；分布式时代，主要解决方案是分布式文件系统。

其实不论是在单机时代还是分布式时代，大规模数据存储都需要解决几个核心问题，这些问题都是什么呢？总结一下，主要有以下三个方面。

1.**数据存储容量的问题。**既然大数据要解决的是数以 PB 计的数据计算问题，而一般的服务器磁盘容量通常 1～2TB，那么如何存储这么大规模的数据呢？

2.**数据读写速度的问题。**一般磁盘的连续读写速度为几十 MB，以这样的速度，几十 PB 的数据恐怕要读写到天荒地老。

3.**数据可靠性的问题。**磁盘大约是计算机设备中最易损坏的硬件了，通常情况一块磁盘使用寿命大概是一年，如果磁盘损坏了，数据怎么办？

在大数据技术出现之前，我们就需要面对这些关于存储的问题，对应的解决方案就是 RAID 技术。今天我们就先从 RAID 开始，一起看看大规模数据存储方式的演化过程。

RAID（独立磁盘冗余阵列）技术是将多块普通磁盘组成一个阵列，共同对外提供服务。主要是为了改善磁盘的存储容量、读写速度，增强磁盘的可用性和容错能力。在 RAID 之前，要使用大容量、高可用、高速访问的存储系统需要专门的存储设备，这类设备价格要比 RAID 的几块普通磁盘贵几十倍。RAID 刚出来的时候给我们的感觉像是一种黑科技，但其原理却不复杂，下面我慢慢道来。

目前服务器级别的计算机都支持插入多块磁盘（8 块或者更多），通过使用 RAID 技术，实现数据在多块磁盘上的并发读写和数据备份。

常用 RAID 技术有图中下面这几种，光看图片你可能觉得它们都差不多，下面我给你讲讲它们之间的区别。

首先，我们先假设服务器有 N 块磁盘，**RAID 0**是数据在从内存缓冲区写入磁盘时，根据磁盘数量将数据分成 N 份，这些数据同时并发写入 N 块磁盘，使得数据整体写入速度是一块磁盘的 N 倍；读取的时候也一样，因此 RAID 0 具有极快的数据读写速度。但是 RAID 0 不做数据备份，N 块磁盘中只要有一块损坏，数据完整性就被破坏，其他磁盘的数据也都无法使用了。

RAID 1是数据在写入磁盘时，将一份数据同时写入两块磁盘，这样任何一块磁盘损坏都不会导致数据丢失，插入一块新磁盘就可以通过复制数据的方式自动修复，具有极高的可靠

性。

结合 RAID 0 和 RAID 1 两种方案构成了**RAID 10**，它是将所有磁盘 N 平均分成两份，数据同时在两份磁盘写入，相当于 RAID 1；但是平分成两份，在每一份磁盘（也就是 N/2 块磁盘）里面，利用 RAID 0 技术并发读写，这样既提高可靠性又改善性能。不过 RAID 10 的磁盘利用率较低，有一半的磁盘用来写备份数据。

一般情况下，一台服务器上很少出现同时损坏两块磁盘的情况，在只损坏一块磁盘的情况下，如果能利用其他磁盘的数据恢复损坏磁盘的数据，这样在保证可靠性和性能的同时，磁盘利用率也得到大幅提升。

顺着这个思路，**RAID 3**可以在数据写入磁盘的时候，将数据分成 N-1 份，并发写入 N-1 块磁盘，并在第 N 块磁盘记录校验数据，这样任何一块磁盘损坏（包括校验数据磁盘），都可以利用其他 N-1 块磁盘的数据修复。

但是在数据修改较多的场景中，任何磁盘数据的修改，都会导致第 N 块磁盘重写校验数据。频繁写入的后果是第 N 块磁盘比其他磁盘更容易损坏，需要频繁更换，所以 RAID 3 很少在实践中使用，因此在上面图中也就没有单独列出。

相比 RAID 3，**RAID 5**是使用更多的方案。RAID 5 和 RAID 3 很相似，但是校验数据不是写入第 N 块磁盘，而是螺旋式地写入所有磁盘中。这样校验数据的修改也被平均到所有磁盘上，避免 RAID 3 频繁写坏一块磁盘的情况。

如果数据需要很高的可靠性，在出现同时损坏两块磁盘的情况下（或者运维管理水平比较落后，坏了一块磁盘但是迟迟没有更换，导致又坏了一块磁盘），仍然需要修复数据，这时候可以使用**RAID 6**。

RAID 6 和 RAID 5 类似，但是数据只写入 N-2 块磁盘，并螺旋式地在两块磁盘中写入校验信息（使用不同算法生成）。

从下面表格中你可以看到在相同磁盘数目（N）的情况下，各种 RAID 技术的比较。

RAID 类型	访问速度	数据可靠性	磁盘利用率
RAID0	很快	很低	100%
RAID1	很慢	很高	50%
RAID10	中等	很高	50%
RAID5	较快	较高	(*N*–1)/*N*
RAID6	较快	较(RAID5)高	(*N*–2)/*N*

RAID 技术有硬件实现，比如专用的 RAID 卡或者主板直接支持；也可以通过软件实现，在操作系统层面将多块磁盘组成 RAID，从逻辑上视作一个访问目录。RAID 技术在传统关系数据库及文件系统中应用比较广泛，是改善计算机存储特性的重要手段。

现在我来总结一下，看看 RAID 是如何解决我一开始提出的，关于存储的三个关键问题。

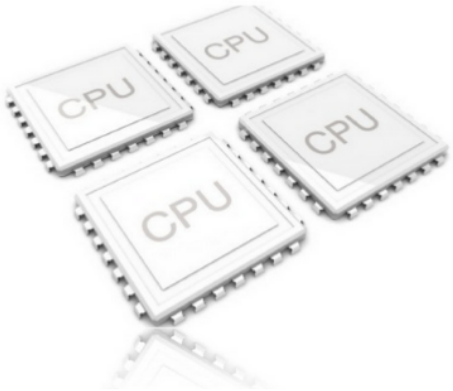
1. 数据存储容量的问题。RAID 使用了 N 块磁盘构成一个存储阵列，如果使用 RAID 5，数据就可以存储在 N-1 块磁盘上，这样将存储空间扩大了 N-1 倍。

2. 数据读写速度的问题。RAID 根据可以使用的磁盘数量，将待写入的数据分成多片，并发同时向多块磁盘进行写入，显然写入的速度可以得到明显提高；同理，读取速度也可以得到明显提高。不过，需要注意的是，由于传统机械磁盘的访问延迟主要来自于寻址时间，数据真正进行读写的时间可能只占据整个数据访问时间的一小部分，所以数据分片后对 N 块磁盘进行并发读写操作并不能将访问速度提高 N 倍。

3. 数据可靠性的问题。使用 RAID 10、RAID 5 或者 RAID 6 方案的时候，由于数据有冗余存储，或者存储校验信息，所以当某块磁盘损坏的时候，可以通过其他磁盘上的数据和校验数据将丢失磁盘上的数据还原。

我们对更强计算能力和更大规模数据存储的追求几乎是没有止境的，这似乎是源于人类的天性。神话里人类试图建立一座通天塔到神居住的地方，就是这种追求的体现。

我在上一期提到过，在计算机领域，实现更强的计算能力和更大规模的数据存储有两种思路，一种是升级计算机，一种是用分布式系统。前一种也被称作“垂直伸缩”（scaling up），通过升级 CPU、内存、磁盘等将一台计算机变得更强大；后一种是“水平伸缩”（scaling out），添加更多的计算机到系统中，从而实现更强大的计算能力。



垂直伸缩 VS 水平伸缩

在计算机发展的早期，我们获得更强大计算能力的手段主要依靠垂直伸缩。一方面拜摩尔定律所赐，每 18 个月计算机的处理能力提升一倍；另一方面由于不断研究新的计算机体系结构，小型机、中型机、大型机、超级计算机，不断刷新我们的认知。

但是到了互联网时代，这种垂直伸缩的路子走不通了，一方面是成本问题，互联网公司面对巨大的不确定性市场，无法为一个潜在的需要巨大计算资源的产品一下投入很多钱去购买大型计算机；另一方面，对于 Google 这样的公司和产品而言，即使是世界上最强大的超级计算机也无法满足其对计算资源的需求。

所以互联网公司走向了一条新的道路：**水平伸缩**，在一个系统中不断添加计算机，以满足不断增长的用户和数据对计算资源的需求。这就是最近十几年引导技术潮流的分布式与大数据技术。

RAID 可以看作是一种垂直伸缩，一台计算机集成更多的磁盘实现数据更大规模、更安全可靠的存储以及更快的访问速度。而 HDFS 则是水平伸缩，通过添加更多的服务器实现数据更大、更快、更安全存储与访问。

RAID 技术只是在单台服务器的多块磁盘上组成阵列，大数据需要更大规模的存储空间和更快的访问速度。将 RAID 思想原理应用到分布式服务器集群上，就形成了 Hadoop 分布式文件系统 HDFS 的架构思想。

垂直伸缩总有尽头，水平伸缩理论上是没有止境的，在实践中，数万台服务器的 HDFS 集群已经出现，我会在下一期谈谈 HDFS 的架构。

思考题

传统机械磁盘进行数据连续写入的时候，比如磁盘以日志格式连续写入操作，其写入速度远远大于磁盘随机写入的速度，比如关系数据库连续更新若干条数据记录，你知道这是为什么吗？

欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

上一篇　04 | 移动计算比移动数据更划算

下一篇　06 | 新技术层出不穷，HDFS依然是存储的王者

精选留言 (69)

写留言

Panmax 置顶

2018-11-08

6

3. 数据可靠性的问题。使用 RAID 0、RAID 5 或者 RAID 6 方案的时候，由于数据有冗余存储，或者存储校验信息，所以当某块磁盘损坏的时候，可以通过其他磁盘上的数据和校

验数据将丢失磁盘上的数据还原。

这里应该是 RAID1 吧

展开

作者回复: 实践中一般用raid10，已订正，谢谢指正

Zach_
2018-11-08

63

连续写入：写入只寻址一次 存储位置与逻辑位置相邻 不用多次寻址

随机写入：每写一次 便寻址一次 增加了磁盘的寻址时间

展开

作者回复: 是的

wmg
2018-11-08

17

到目前为止专栏的内容基本上是普及大数据知识，非常适合打算入坑的码农，期待后续能有更多关于大数据系统架构和针对某项技术深入介绍的内容。

o°cboy
2018-11-08

8

磁盘的读写过程，最消耗时间的地方就是在磁盘中磁道寻址的过程，而一旦寻址完成，写入数据的速度很快。
顺序写入只要一次寻址操作，而随机写入要多次寻址操作。所以顺序写入速度明显高于随机写入。
个人的理解，不正确的地方，还请多多指教。

展开

作者回复: 是的

Zach_
2018-11-08

7

老师居然回我信息了，好开心！ 我最喜欢那种 讲课做事都亲自来的老师！ 听了老师四节课了，都是老师自己读，有的话是老师的原汁原味的话，在文稿里没有！ 给智慧老师打call!

展开

作者回复: 谢谢

lyshrine
2018-11-08

7

老师，为啥通常情况一块磁盘使用寿命大概是一年？磁盘不是能用很多年吗？一年一换成本会不会太高了？

作者回复: 服务器磁盘访问压力大，寿命短
你的电脑常年不关机下小电影，硬盘也坏的快

hashmap
2018-11-09

5

磁盘寻址是耗时操作，是时间大于写入时间
连续写入，可以寻址一次，然后写入
随机写入，需要寻址多次，然后写入
所以连续写入快
这个问题可以延伸回答，为什么很多数据库索引采用b+树，而不是完全二叉树？...

展开

作者回复: 有raid硬件，也有驱动实现

暴风雪
2018-11-09

4

那个RAID3的修复，可以理解为：b1+b2+b3+...+bn=s，其中一块坏掉了，也就是bn数据不见了，可以通过bn=s-b1-b2-b3-...-b(n-1)。

落叶飞逝的...
2018-11-08

4

RAID 5 6螺旋写入这个怎么看？前面三个图的DATA表示看的懂，后面两个看不懂？还有就是平常开发接触不到服务器怎么办？

格非
2018-11-08

4

跟机械磁盘的构造有关，随机读写时，磁头需要不停的移动，时间都浪费在了磁头寻址上

作者回复: 是的

qpm
2018-11-08

3

传统机械硬盘的读写耗时主要在寻址上，连续读写一般只寻址一次，所以速度会快。

作者回复: 是的

暴风雪
2018-11-09

2

思考题：上文说过，机械硬盘时间消耗主要在寻址上，所以我猜想，连续文件写入时是只寻址一次，后面可以连续写入，所以时间计划不变，而数据库随机写入，每次都要寻址，分配新的地址，所以时间就慢很多了

展开

zc
2018-11-08

2

老师请推荐大数据相关书籍

展开

lyshrine
2018-11-08

2

“RAID 3可以在数据写入磁盘的时候，将数据分成 N-1 份，并发写入 N-1 块磁盘，并在第 N 块磁盘记录校验数据，这样任何一块磁盘损坏（包括校验数据磁盘），都可以利用其他 N-1 块磁盘的数据修复。”
不是很明白：数据都是写到N-1的磁盘里，每个磁盘里的数据都不一样，没有备份，如何数据修复?

展开

公号-代码...
2018-11-08

2

1 计算写入地址更简单快速
2 磁盘机械机构移动的距离更少，寻址更快
3 由于空间的连续性，写入也更快

展开

作者回复: 是的

刘工的一号...
2018-11-29

1

RAID5为什么是N-1呢？不是所有磁盘螺旋写入吗？应该所有磁盘都可以使用啊

GeXeLr
2018-11-16

1

老师那个磁盘利用率是怎么计算出来的呀？还有速度提升倍数又是怎么计算出来的？

沙鱼
2018-11-13

1

老师请问一下：如果部署hdfs集群，还有没有必要单台机器做raid5浪费磁盘呢？想听听老师的分析。

作者回复: 没有，hdfs要保证机架级别的数据可用性，raid5解决不了，请看下期。

暴风雪
2018-11-09

1

我想问一下，RAID 3的任意一块磁盘损坏，通过其他磁盘的数据修复，是怎么修复的？有点不理解这段话

作者回复: 有一块盘记录校验数据，用校验数据和未损坏盘数据可以计算损坏盘的数据

才才
2018-11-08

1

批量地址连续，指针的寻址没那么跳跃

展开

作者回复: 是的