第一章

硬件

经常有人问我： 做数据挖掘时最佳硬件配置是什么？对于这类型的问题， 唯一合适的答案应该是：这取决于你想要做什么。 在决定如何为大数据分析构建适当的计算环境时，需要考虑许多因素。

储存（磁盘）

当提及大数据时，数据存储通常是第一个被想到的。数据存储允许我们保留历史记录，以便它可以告诉我们将来可能发生的事情。

传统的硬盘驱动器由盘片组成，这些盘片外覆盖有磁铁薄片，它可以用来记录由1和0编成的数据。转动垂直堆叠的盘片的主轴是评级硬盘驱动器的关键部分，因为主轴可以确定盘片旋转的速度，从而确定读取和写入数据的速度。每个盘片都有一个驱动头; 它们都一致地移动，以便只有一个驱动头从特定盘片中读取。

为了克服磁盘速度的弱点，磁盘阵列[[1]](#footnote-1)变得广泛可用，并且它们提供更高的吞吐量。从外部存储子系统到单个系统的磁盘阵列的最大吞吐量在1到6千兆字节（GB）/秒 范围内（数据访问速率的10到50倍的加速）。

作为对大数据时代的回应，磁盘驱动器的另一个变化是，在过去10年中，它们的容量每年增加50％到100％。 此外，磁盘阵列的价格几乎保持不变，也就是说每TB的价格每年下降一半。

磁盘驱动器每年仅增加15％到20％，这意味着磁盘驱动器容量的增加与磁盘驱动器之间的数据传输能力并不匹配。比如说，在2008年，典型的服务器驱动器为500 GB，数据传输速率为每秒98兆字节（MB /秒）。 整个磁盘可以在大约85分钟内传输（500 GB = 500,000 MB / 98 MB /秒）。 在2013年，有4 TB磁盘的传输速率为150 MB /秒，但传输整个磁盘大约需要440分钟。考虑到每隔几年翻一番的数据量，这个问题就很明显了。我们需要更快的磁盘。

固态设备（SSD）是没有磁盘或任何移动部件的磁盘驱动器。 它们可以被认为是稳定的内存，它们的数据读取速率可以轻松超过450 MB /秒。 对于中等规模的数据挖掘环境，SSD及其卓越的吞吐率可以显着改变解决方案的时间。 SSD阵列也可用，但SSD的单位容量成本仍然比硬盘驱动器（HDD）高得多。 SSD阵列受到与HDD阵列相同的外部存储带宽的限制。 因此，尽管SSD可以通过减少读取和写入数据的总时间来解决数据挖掘问题，但将所有存储转换为SSD可能会成本过高。 在这种情况下，需要使用不同类型设备的混合策略。

另一个考虑因素是为分析工作负载购买的磁盘驱动器的大小。 较小的磁盘具有更快的访问时间，并且来自多个磁盘的并行磁盘访问可以同时为同一问题读取数据。 仅当软件可以利用此类磁盘驱动器配置时，这才是一个优势。

从历史上看，只有一些分析软件能够通过将中间结果写入磁盘存储来使用额外的存储来增加内存。这扩大了可以解决的问题的大小，但导致运行时间上升。运行时间上升不仅是因为额外的数据加载，而且还因为它不是从内存中读取中间结果而进行的是速度较慢的从磁盘读取中间结果。对于典型的桌面或小型服务器系统，对存储设备的数据访问（尤其是写入存储设备）非常缓慢。分析过程的单个执行线程可以轻松消耗100 MB /秒，主要类型的数据访问是顺序读取或写入。 典型的高端工作站具有15K RPM SAS驱动器; 驱动器以每分钟15,000转的速度旋转，并使用SAS技术以100至150 MB /秒的速率读取和写入数据。 这意味着一个或两个核心可以消耗所有可用的磁盘带宽。 这也意味着在具有许多内核的现代系统中，用于许多数据挖掘活动的很大一部分中央处理单元（CPU）资源将闲置; 这不是缺少所需的计算资源，而是磁盘，内存和CPU之间存在的不匹配。

中央处理器

术语“CPU”在计算机硬件中有两个含义。 CPU用于指塑料和钢制外壳，它包含计算机的所有基本元素。 这包括电源，主板，外围设备卡等。 CPU的另一个含义是位于塑料和钢箱内的加工芯片。 在本书中，CPU指的是芯片。

在20世纪80年代和90年代，CPU的速度显着提高。 CPU速度正在以这样的速度增长：单线程软件应用程序在新的CPU版本上运行时的速度几乎是可用的两倍。英特尔的联合创始人戈登•摩尔（Gordon Moore）在著名的摩尔定律中描述了CPU加速，这是一个观察结果，即能够在给定区域内放置的晶体管和集成电路的数量每两年增加一倍，因此指令可以是以两倍的速度执行。这种将CPU速度提高一倍的趋势一直持续到20世纪90年代，当时英特尔工程师观察到，如果双倍的趋势继续下去，那么从这些芯片发出的热量将在2010年与太阳一样热。在21世纪初，摩尔定律免费午餐已经结束，至少在处理速度方面。 处理器速度（频率）停滞不前，计算机公司寻求提高性能的新方法。 自Pentium MMX指令以来在x86中以有限形式呈现的向量单元对于获得性能和获得额外功能（例如单精度和双精度oating点）越来越重要。

在21世纪初期，芯片制造商也开始在芯片中添加额外的执行线程。 这些多核芯片是多处理器超级计算机的缩小版本，有着核心共享诸如高速缓冲存储器之类的资源。位于单个芯片上的核心数量随着时间的推移而增加; 今天许多服务器机器提供两个六核CPU。

与硬盘数据访问相比，CPU对内存的访问速度快于子弹; 典型的访问范围为10到30 GB /秒。 计算机的所有其他组件都在赛跑一般以跟上CPU的步伐。

图形处理单元

图形处理单元（GPU）作为未使用的计算资源得到了相当大的宣传，通过并行化计算可以减少数据挖掘和其他分析问题的运行时间。 GPU存在于世界上的每台台式计算机中。

在2000年代早期，GPU进入了计算游戏。 图形处理已从第一台台式计算机的早期纯文本显示中大大改进。行业对可视化工具的需求推动着对更好图形的追求。举个例子，在构建新设计之前，工程师使用三维（3D）计算机辅助设计（CAD）软件创建新设计的原型。消费者视频游戏行业是GPU计算的一个更大的驱动力，其价格和性能趋势与消费者计算行业的其他部分类似。 以更低的成本不断推动更高性能的发展使得普通用户在CPU和GPU上都获得了前所未有的性能。

在创建动画时，三维图形处理必须每秒多次处理3D场景中的数百万或数十亿个3D三角形。在3D环境中放置和着色所有这些三角形需要大量非常相似的计算。最初，3D图形最初是通过使用固定渲染管道完成的，该管道获取3D场景信息并将其转换为可在视频或屏幕上呈现给用户的像素。这个固定的管道是存在于硬件中的，与GPU的各个不同的部分一起在将三角形转换为像素的问题上做着不同的工作。在21世纪初，这种固定的管道逐渐被通用的软件着色器所取代，这些着色器是执行早期固定硬件管道操作的小型程序。

通过这些着色器，高性能计算人员注意到，如果你恰到好处地观察它们，浮点坐标和颜色可能看起来很像物理或化学问题。更多的硬件底层的更改创建出图形问题，这个看起来很没道理，除了基础计算完成解决难题非常快。性能提升得到了注意，并且开发了使用GPU进行非图形计算的计算框架。 这些计算与数据挖掘所需的类型相同。

GPU是一个绿色领域。 从历史上看，在GPU上开发和运行代码的能力是有限的并且代价高昂。 在过去几年中，利用GPU开发软件的编程接口有了很大的改进。 软件利用GPU只是个开始，而将数据挖掘所需的计算委托给GPU执行还需要几年的时间。 到那个时候，许多类型的数据挖掘问题的加速将从几小时减少到几分钟，从几分钟减少到几分钟。

内存

通常所称的内存或随机存取存储器（RAM）是构建数据挖掘平台的关键且经常被低估的组件。 内存是数据存储和CPU执行的数学运算处理之间的中介。 内存是易失性的，这意味着如果它断电，存储在其中的数据就会丢失。

在20世纪80年代和90年代，数据挖掘算法的发展受到内存和CPU的限制。内存限制是由32位操作系统引起的，它只允许处理4 GB的内存。 这个限制实际上意味着超过4 GB的内存[[2]](#footnote-2)（减去机器上运行的软件和操作系统）的数据挖掘问题不可以单独使用内存。 这非常重要，因为内存的数据吞吐量通常为12到30 GB /秒，而最快的存储只有大约6 GB /秒，而大多数存储吞吐量要低得多。

2004年左右，商用硬件（英特尔和AMD）支持64位计算。 与此同时，操作系统变得能够支持更大量的内存，实际内存价格急剧下降。 2000年，1 MB RAM的平均价格为1.12美元。 2005年，平均价格为0.185美元; 而在2010年，它是0.0122美元。

有了64位计算系统的支持，可以解决高达8 TB的内存并且内存价格有所下降，现在可以构建可以将整个数据挖掘问题存储在内存中的数据挖掘平台。这在很短的时间内就可以产生结果。

数据挖掘算法通常要求所有数据和计算都在内存中完成。 在没有外部存储的情况下，虚拟和实际地址空间的增加以及内存价格的急剧下降为解决以前不可行的许多数据挖掘问题创造了机会。

为了说明此示例，请考虑使用神经网络算法的预测建模问题。 神经网络将执行迭代优化以找到最佳模型。 对于每次迭代，它必须一次读取数据。神经网络通过数据进行数千次传递以找到最佳解决方案并不罕见。 如果这些传递在内存中以20 GB /秒的速度完成，而在磁盘上以1 GB /秒的速度完成，则内存中仅需10秒解决的问题用磁盘解决会超过3分钟。如果经常重复这种情况，数据采矿者的生产率会急剧下降。 除了人力资本的生产力之外，如果数据挖掘过程依赖于磁盘存储，则计算将花费很多倍来完成。完成一个过程所需的时间越长，某种硬件故障的可能性就越高。 这些类型的故障通常是不可恢复的，并且必须重新启动整个过程。内存速度以比处理器速度更加适中的速度增加。 与处理器速度相比，内存速度提高了10倍，处理器速度提高了10,000倍。磁盘存储吞吐量的增长速度甚至比内存还要慢。 因此，数据挖掘算法主要维护内存中的所有数据结构，并已转向分布式计算以增加计算和内存容量。内存带宽通常在12到30 GB /秒的范围内，内存非常便宜。 高带宽存储最大范围为6 GB /秒，非常昂贵。 与购买昂贵的高速磁盘存储系统相比，部署具有大量内存的商用系统要便宜得多。

今天的现代服务器系统通常装载64 GB和256 GB的内存。 要获得快速结果，必须考虑内存的大小。

网络

网络是计算机外部[[3]](#footnote-3)唯一的硬件组件.它是计算机与其他计算机通信的机制。 除了本节非常有限的细节之外，这里将不讨论网络通信的许多协议和标准。网络速度应该只是分布式计算环境的一个因素。 对于单台计算机（工作站或服务器），数据，内存和CPU都应该是本地的，分析任务的性能不受网络速度的影响。

分析计算集群的标准网络连接是10千兆以太网（10 GbE），其上限数据传输速率为每秒4千兆字节（GB /秒）。 这种数据传输速率远远低于已经讨论过的任何其他基本要素。 Inniband®等专有协议可提供更好的数据吞吐量，但仍与其他组件的速度不匹配。 出于这个原因，最小化网络对数据移动的使用以及甚至计算设备中的不同节点之间的非必要通信是非常重要的。

正是这种网络速度瓶颈使得许多数据挖掘算法的并行化变得如此具有挑战性。 在不移动任何数据和限制计算机之间的通信的同时，需要使用众多算法中的一种来有效和精确地使用软件基础结构，算法选择和最终实现方面的大量技能。

如果您拥有分布式计算环境，则高性能数据挖掘平台的网络速度将非常重要。 由于网络数据传输速率比其他组件慢得多，因此在评估数据挖掘软件解决方案时必须考虑网络组件。

1. 磁盘阵列是一种专用硬件存储，由于其专门的实现，可提供更大的存储容量和数据访问。 NetApp和EMC是磁盘阵列的两个主要供应商。 [↑](#footnote-ref-1)
2. 32位元操作系统可用于处理或引用内存的最大整数值为232-1或3.73 GB。 [↑](#footnote-ref-2)
3. 存储有时可能是存储区域网络（SAN）中的外部存储。 [↑](#footnote-ref-3)