PART
ONE

The Computing Environment

With data collection, "the sooner the better" is always the best answer.

— Marissa Mayer

Data mining is going through a significant shift with the volume, variety, value and velocity of data increasing significantly each year. The volume of data created is outpacing the amount of currently usable data to such a degree that most organizations do not know what value is in their data. At the same that time data mining is changing, hardware capabilities have also undergone dramatic changes. Just as data mining is not one thing but a collection of many steps, theories, and algorithms, hardware can be dissected into a number of components. The corresponding component changes are not always in sync with this increased demand in data mining, machine learning, and big analytical problems.

The four components of disk, memory, central processing unit, and network can be thought of as four legs of the hardware platform stool. To have a useful stool, all the legs must be of the same length or users will be frustrated, stand up, and walk away to find a better stool; so too must the hardware system for data mining be in balance in regard to the components to give users the best experience for their analytical problems.

Data mining on any scale cannot be done without specialized software. In order to explain the evolution and progression of the hardware, there needs to be a small amount of background on the traditional interaction between hardware and software. Data mining software packages are discussed in detail in Part One.

In the past, traditional data mining software was implemented by loading data into memory and running a single thread of execution over the data. The process was constrained by the amount of memory available and the speed of a processor. If the process could not fit entirely into memory, the process would fail. The single thread of execution also failed to take advantage of multicore servers unless multiple users were on the system at the same time.

The main reason we are seeing dramatic changes in data mining is related to the changes in storage technologies as well as computational capabilities. However, all software packages cannot take advantage of current hardware capacity. This is especially true of the distributed computing model. A careful evaluation should be made to ensure that algorithms are distributed and effectively leveraging all the computing power available to you.

CHAPTER 1

Hardware

I am often asked what the best hardware configuration is for doing data mining. The only appropriate answer for this type of question is that it depends on what you are trying to do. There are a number of considerations to be weighed when deciding how to build an appropriate computing environment for your big data analytics.

STORAGE (DISK)

Storage of data is usually the first thing that comes to mind when the topic of big data is mentioned. It is the storage of data that allows us to keep a record of history so that it can be used to tell us what will likely happen in the future.

A traditional hard drive is made up of platters which are actual disks coated in a magnetized film that allow the encoding of 1s and 0s that make up data. The spindles that turn the vertically stacked platters are a critical part of rating hard drives because the spindles determine how fast the platters can spin and thus how fast the data can be read and written. Each platter has a single drive head; they both move in unison so that only one drive head is reading from a particular platter.

This mechanical operation is very precise and also very slow compared to the other components of the computer. It can be a large

contributor to the time required to solve high-performance data mining problems.

To combat the weakness of disk speeds, disk arrays[1] became widely available, and they provide higher throughput. The maximum throughput of a disk array to a single system from external storage subsystems is in the range of 1 to 6 gigabytes (GB) per second (a speedup of 10 to 50 times in data access rates).

Another change in disk drives as a response to the big data era is that their capacity has increased 50% to 100% per year in the last 10 years. In addition, prices for disk arrays have remained nearly constant, which means the price per terabyte (TB) has decreased by half per year.

This increase in disk drive capacity has not been matched by the ability to transfer data to/from the disk drive, which has increased by only 15% to 20% per year. To illustrate this, in 2008, the typical server drive was 500 GB and had a data transfer rate of 98 megabytes per second (MB/sec). The entire disk could be transferred in about 85 minutes (500 GB = 500,000 MB/98 MB/sec). In 2013, there were 4 TB disks that have a transfer rate of 150 MB/sec, but it would take about 440 minutes to transfer the entire disk. When this is considered in light of the amount of data doubling every few years, the problem is obvious. Faster disks are needed.

Solid state devices (SSDs) are disk drives without a disk or any moving parts. They can be thought of as stable memory, and their data read rates can easily exceed 450 MB/sec. For moderate-size data mining environments, SSDs and their superior throughput rates can dramatically change the time to solution. SSD arrays are also available, but SSDs still cost significantly more per unit of capacity than hard disk drives (HDDs). SSD arrays are limited by the same external storage bandwidth as HDD arrays. So although SSDs can solve the data mining problem by reducing the overall time to read and write the data, converting all storage to SSD might be cost prohibitive. In this case, hybrid strategies that use different types of devices are needed.

Another consideration is the size of disk drives that are purchased for analytical workloads. Smaller disks have faster access times, and

[1] A disk array is a specialized hardware storage that provides larger storage capacity and data access because of its specialized implementation. NetApp and EMC are two major vendors of disk arrays.

there can be advantages in the parallel disk access that comes from multiple disks reading data at the same time for the same problem. This is an advantage only if the software can take advantage of this type of disk drive configuration.

Historically, only some analytical software was capable of using additional storage to augment memory by writing intermediate results to disk storage. This extended the size of problem that could be solved but caused run times to go up. Run times rose not just because of the additional data load but also due to the slower access of reading intermediate results from disk instead of reading them from memory. For a typical desktop or small server system, data access to storage devices, particularly writing to storage devices, is painfully slow. A single thread of execution for an analytic process can easily consume 100 MB/sec, and the dominant type of data access is sequential read or write. A typical high-end workstation has a15K RPM SAS drive; the drive spins at 15,000 revolutions per minute and uses the SAS technology to read and write data at a rate of 100 to 150 MB/sec. This means that one or two cores can consume all of the disk bandwidth available. It also means that on a modern system with many cores, a large percentage of the central processing unit (CPU) resources will be idle for many data mining activities; this is not a lack of needed computation resources but the mismatch that exists among disk, memory, and CPU.

CENTRAL PROCESSING UNIT

The term "CPU" has had two meanings in computer hardware. CPU is used to refer to the plastic and steel case that holds all the essential elements of a computer. This includes the power supply, motherboard, peripheral cards, and so on. The other meaning of CPU is the processing chip located inside the plastic and steel box. In this book, CPU refers to the chip.

The speed of the CPU saw dramatic improvements in the 1980s and 1990s. CPU speed was increasing at such a rate that single threaded software applications would run almost twice as fast on new CPU versions as they became available. The CPU speedup was described by Gordon Moore, cofounder of Intel, in the famous Moore's law, which is an observation that the number of transistors and integrated circuits

that are able to be put in a given area doubles every two years and therefore instructions can be executed at twice the speed. This trend in doubling CPU speed continued into the 1990s, when Intel engineers observed that if the doubling trend continued, the heat that would be emitted from these chips would be as hot as the sun by 2010. In the early 2000s, the Moore's law free lunch was over, at least in terms of processing speed. Processor speeds (frequencies) stalled, and computer companies sought new ways to increase performance. Vector units, present in limited form in x86 since the Pentium MMX instructions, were increasingly important to attaining performance and gained additional features, such as single- and then double-precision floating point.

In the early 2000s, then, chip manufacturers also turned to adding extra threads of execution into their chips. These multicore chips were scaled-down versions of the multiprocessor supercomputers, with the cores sharing resources such as cache memory. The number of cores located on a single chip has increased over time; today many server machines offer two six-core CPUs.

In comparison to hard disk data access, CPU access to memory is faster than a speeding bullet; the typical access is in the range of 10 to 30 GB/sec. All other components of the computer are racing to keep up with the CPU.

Graphical Processing Unit

The graphical processing unit (GPU) has gotten considerable publicity as an unused computing resource that could reduce the run times of data mining and other analytical problems by parallelizing the computations. The GPU is already found in every desktop computer in the world.

In the early 2000s, GPUs got into the computing game. Graphics processing has evolved considerably from early text-only displays of the first desktop computers. This quest for better graphics has been driven by industry needs for visualization tools. One example is engineers using three-dimensional (3D) computer-aided design (CAD) software to create prototypes of new designs prior to ever building them. An even bigger driver of GPU computing has been the consumer video game industry, which has seen price and performance trends similar to the rest of the consumer computing industry. The relentless

drive to higher performance at lower cost has given the average user unheard-of performance both on the CPU and the GPU.

Three-dimensional graphics processing must process millions or billions of 3D triangles in 3D scenes multiple times per second to create animation. Placing and coloring all of these triangles in their 3D environment requires a huge number of very similar calculations. Initially, 3D graphics were done using a fixed rendering pipeline, which took the 3D scene information and turned it into pixels that could be presented to the user in a video or on the screen. This fixed pipeline was implemented in hardware, with various parts of the GPU doing different pieces of the problem of turning triangles into pixels. In the early 2000s, this fixed pipeline was gradually replaced by generalized software shaders, which were miniprograms that performed the operations of the earlier fixed hardware pipeline.

With these shaders, high-performance computing folks noticed that the floating-point coordinates and colors could look an awful lot like physics or chemistry problems if you looked at them just right. The more hardcore hacker types started creating graphics problems that looked a lot like nonsense except that the underlying calculations being done solved hard problems remarkably fast. The performance gains got noticed, and computing frameworks, which used the GPUs for doing nongraphics calculations, were developed. These calculations are the same type needed for data mining.

GPUs are a green field. Historically the ability to develop code to run on the GPU was restrictive and costly. Those programming interfaces for developing software that takes advantage of GPUs have improved greatly in the last few years. Software has only started to take advantage of the GPU, and it will be several years before the computations needed for data mining are efficiently delegated to the GPU for execution. When that time comes, the speedup in many types of data mining problems will be reduced from hours to minutes and from minutes to seconds.

MEMORY

Memory, or random access memory (RAM) as it is commonly referred to, is the crucial and often undervalued component in building a data mining platform. Memory is the intermediary between the storage of

data and the processing of mathematical operations that are performed by the CPU. Memory is volatile, which means that if it loses power, the data stored in it is lost.

In the 1980s and 1990s, the development of data mining algorithms was very constrained by both memory and CPU. The memory constraint was due to the 32-bit operating systems, which allow only 4 GB of memory to be addressed. This limit effectively meant that no data mining problem that required more than 4 GB of memory[2] (minus the software and operating system running on the machine) could be done using memory alone. This is very significant because the data throughput of memory is typically 12 to 30 GB/sec, and the fastest storage is only around 6 GB/sec with most storage throughput being much less.

Around 2004, commodity hardware (Intel and AMD) supported 64-bit computing. At the same time operating systems became capable of supporting larger amounts of memory, the actual price of memory dropped dramatically. In 2000, the average price of 1 MB of RAM was $1.12. In 2005, the average price was $0.185; and in 2010, it was $0.0122.

With this support of 64-bit computing systems that can address up to 8 TB of memory and the drop in memory prices, it was now possible to build data mining platforms that could store the entire data mining problem in memory. This in turn produced results in a fraction of the time.

Data mining algorithms often require all data and computation to be done in memory. Without external storage, the increase in virtual and real address space as well as the dramatic drop in the price of memory created an opportunity to solve many data mining problems that previously were not feasible.

To illustrate this example, consider a predictive modeling problem that uses a neural network algorithm. The neural network will perform an iterative optimization to find the best model. For each iteration, it will have to read the data one time. It is not uncommon for neural networks to make thousands of passes through the data to find

[2] The largest integer value that 32-bit operating systems can use to address or reference memory is 2^{32}–1, or 3.73 GB, of memory.

the optimal solution. If these passes are done in memory at 20 GB/sec versus on disk at 1 GB/sec, a problem that is only 10 seconds to solve in memory will be more than 3 minutes to solve using disk. If this scenario is repeated often, the productivity of the data miner plummets. In addition to the productivity of the human capital, if the data mining processes relied on disk storage, the computation would take many times longer to complete. The longer a process takes to complete, the higher the probability of some sort of hardware failure. These types of failure are typically unrecoverable, and the entire process must be restarted.

Memory speeds have increased at a much more moderate rate than processor speeds. Memory speeds have increased by 10 times compared to processor speeds, which have increased 10,000 times. Disk storage throughput has been growing at an even slower rate than memory. As a result, data mining algorithms predominantly maintain all data structures in memory and have moved to distributed computing to increase both computation and memory capacity. Memory bandwidth is typically in the 12 to 30 GB/sec range, and memory is very inexpensive. High-bandwidth storage maxes out in the 6 GB/sec range and is extremely expensive. It is much less expensive to deploy a set of commodity systems with healthy amounts of memory than to purchase expensive high-speed disk storage systems.

Today's modern server systems typically come loaded with between 64 GB and 256 GB of memory. To get fast results, the sizing of memory must be considered.

NETWORK

The network is the only hardware component that is always external to the computer.[3] It is the mechanism for computers to communicate to other computers. The many protocols and standards for network communication will not be discussed here beyond the very limited details in this section.

The network speed should be a factor only for a distributed computing environment. In the case of a single computer (workstation or

[3] Storage can sometimes be external in a storage area network (SAN).

server), the data, memory, and CPU should all be local, and performance of your analytical task will be unaffected by network speeds.

The standard network connection for an analytical computing cluster is 10 gigabit Ethernet (10 GbE), which has an upper-bound data transfer rate of 4 gigabytes per second (GB/sec). This data transfer rate is far slower than any of the other essential elements that have been discussed. Proprietary protocols like Infiniband® give better data throughput but still do not match the speed of the other components. For this reason, it is very important to minimize usage of the network for data movement or even nonessential communication between the different nodes in the computing appliance.

It is this network speed bottleneck that makes parallelization of a number of the data mining algorithms so challenging. Considerable skill in the software infrastructure, algorithm selection, and final implementation is required to fit a model efficiently and precisely using one of many algorithms while not moving any data and limiting communication between computers.

The network speed of your high-performance data mining platform will be important if you have a distributed computing environment. Because the network data transfer rate is much slower than other components, you must consider the network component when evaluating data mining software solutions.