Chapter 5

Virtualization and How It Supports Distributed Computing

In This Chapter

- Defining virtualization
- Understanding the hypervisor
- Exploring abstraction and virtualization
- Implementing virtualization to work with big data

Virtualization is a foundational technology applicable to the implementation of both cloud computing and big data. It provides the basis for many of the platform attributes required to access, store, analyze, and manage the distributed computing components in big data environments. Virtualization — the process of using computer resources to imitate other resources — is valued for its capability to increase IT resource utilization, efficiency, and scalability. One primary application of virtualization is server consolidation, which helps organizations increase the utilization of physical servers and potentially save on infrastructure costs. However, you find many benefits to virtualization. Companies that initially focused solely on server virtualization are now recognizing that it can be applied across the entire IT infrastructure, including software, storage, and networks.

In this chapter, we define virtualization and provide insight into the benefits and challenges of virtualized environments. Our primary focus is on the role of virtualization in big data.

Understanding the Basics of Virtualization

Virtualization separates resources and services from the underlying physical delivery environment, enabling you to create many virtual systems within a single physical system. Figure 5-1 shows a typical virtualization environment.

One of the primary reasons that companies have implemented virtualization is to improve the performance and efficiency of processing of a diverse mix of workloads. Rather than assigning a dedicated set of physical resources to each set of tasks, a pooled set of virtual resources can be quickly allocated as needed across all workloads. Reliance on the pool of virtual resources allows companies to improve latency. This increase in service delivery speed and efficiency is a function of the distributed nature of virtualized environments and helps to improve overall time-to-value.

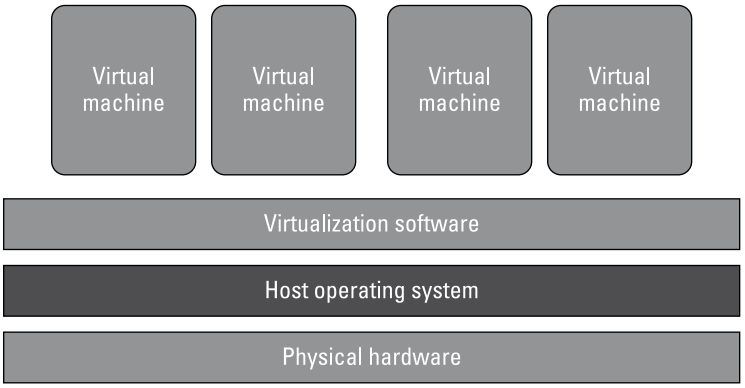

Figure 5-1: Using virtualization software to create several virtual systems within a single physical system.

Using a distributed set of physical resources, such as servers, in a more flexible and efficient way delivers significant benefits in terms of cost savings and improvements in productivity. The practice has several benefits, including the following:

- Virtualization of physical resources (such as servers, storage, and networks) enables substantial improvement in the utilization of these resources.
- Virtualization enables improved control over the usage and performance of your IT resources.
- Virtualization can provide a level of automation and standardization to optimize your computing environment.
- Virtualization provides a foundation for cloud computing.

Although being able to virtualize resources adds a huge amount of efficiency, it doesn't come without a cost. Virtual resources have to be managed so that they are secure. An image can become a technique for an intruder to get direct access to critical systems. In addition, if companies do not have a process for deleting unused images, systems will no longer behave efficiently.

The importance of virtualization to big data

Solving big data challenges typically requires the management of large volumes of highly distributed data stores along with the use of compute- and data-intensive applications. Therefore, you need a highly efficient IT environment to support big data. Virtualization provides the added level of efficiency to make big data platforms a reality. Although virtualization is technically not a requirement for big data analysis, software frameworks such as MapReduce, which are used in big data environments, are more efficient in a virtualized environment. Chapter 8 covers MapReduce in more detail. If you need your big data environment to scale — almost without bounds — you should virtualize elements of your environment.

Virtualization has three characteristics that support the scalability and operating efficiency required for big data environments:

- **Partitioning:** In virtualization, many applications and operating systems are supported in a single physical system by partitioning (separating) the available resources.
- **Isolation:** Each virtual machine is isolated from its host physical system and other virtualized machines. Because of this isolation, if one virtual instance crashes, the other virtual machines and the host system aren't affected. In addition, data isn't shared between one virtual instance and another.
- **Encapsulation:** A virtual machine can be represented (and even stored) as a single file, so you can identify it easily based on the services it provides. For example, the file containing the encapsulated process could be a complete business service. This encapsulated virtual machine could be presented to an application as a complete entity. Thus, encapsulation could protect each application so that it doesn't interfere with another application.

One of the most important requirements for success with big data is having the right level of performance to support the analysis of large volumes and varied types of data. As you begin to leverage environments such as Hadoop and MapReduce, it is critical that you have a supporting infrastructure that can scale. Virtualization adds efficiency at every layer of the IT infrastructure. Applying virtualization across your environment will help to achieve the scalability required for big data analysis.

Implementing virtualization by following an end-to-end approach will deliver benefits for big data and other types of workloads in your environment. An

end-to-end approach will mean that errors can be corrected more quickly — a requirement in a big data environment. When working with big data, your infrastructure needs to be prepared to manage data that is potentially very large (volume), very fast (velocity), and highly unstructured (variety).

As a result, your entire IT environment needs to be optimized at every layer, from the network to the databases, storage, and servers. If you only virtualize your servers, you may experience bottlenecks from other infrastructure elements such as storage and networks. If you only focus on virtualizing one element of your infrastructure, you are less likely to achieve the latency and efficiency you need and more likely to expose your company to higher costs and security risks.

The reality is that most organizations do not attempt to virtualize all elements of their infrastructures at one time. Many organizations begin with server virtualization and achieve a certain level of efficiency improvements. Realistically, other elements may be virtualized as needed to continue to improve overall system performance and efficiency. The following describes how virtualization of each element across the IT environment — servers, storage, applications, data, networks, processors, memory, and services — can have a positive impact on big data analysis.

Server virtualization

In server virtualization, one physical server is partitioned into multiple virtual servers. The hardware and resources of a machine — including the random access memory (RAM), CPU, hard drive, and network controller — can be virtualized (logically split) into a series of virtual machines that each runs its own applications and operating system. A virtual machine (VM) is a software representation of a physical machine that can execute or perform the same functions as the physical machine. A thin layer of software is actually inserted into the hardware that contains a virtual machine monitor, or *hypervisor.* The hypervisor can be thought of as the technology that manages traffic between the VMs and the physical machine.

Server virtualization uses the hypervisor to provide efficiency in the use of physical resources. Of course, installation, configuration, and administrative tasks are associated with setting up these virtual machines. This includes license management, network management, and workload administration, as well as capacity planning.

Server virtualization helps to ensure that your platform can scale as needed to handle the large volumes and varied types of data included in your big data analysis. You may not know the extent of the volume or variety of structured and unstructured data needed before you begin your analysis. This

uncertainty makes the need for server virtualization even greater, providing your environment with the capability to meet the unanticipated demand for processing very large data sets.

In addition, server virtualization provides the foundation that enables many of the cloud services used as data sources in a big data analysis. Virtualization increases the efficiency of the cloud that makes many complex systems easier to optimize. As a result, organizations have the performance and optimization to be able to access data that was previously either unavailable or very hard to collect. Big data platforms are increasingly used as sources of enormous amounts of data about customer preferences, sentiment, and behaviors. Companies can integrate this information with internal sales and product data to gain insight into customer preferences to make more targeted and personalized offers.

Application virtualization

Application infrastructure virtualization provides an efficient way to manage applications in context with customer demand. The application is encapsulated in a way that removes its dependencies from the underlying physical computer system. This helps to improve the overall manageability and portability of the application. In addition, the application infrastructure virtualization software typically allows for codifying business and technical usage policies to make sure that each of your applications leverages virtual and physical resources in a predictable way. Efficiencies are gained because you can more easily distribute IT resources according to the relative business value of your applications. In other words, your most critical applications can receive top priority to draw from pools of available computing and storage capacity as needed.

Application infrastructure virtualization used in combination with server virtualization can help to ensure that business service-level agreements (SLAs) are met. Server virtualization monitors CPU and memory usage, but does not account for variations in business priority when allocating resources. For example, you might require that all applications are treated with the same business-level priority. By implementing application infrastructure virtualization in addition to server virtualization, you can ensure that the most high-priority applications have top-priority access to resources.

Your big data applications may have significant IT resource requirements, due to the large volumes of data or the speed at which that data is generated. Your big data environment needs to have the right level of predictability and repeatability to make sure that the applications have access to the required resources. Application infrastructure virtualization can ensure that each application deployed for a big data analysis has access to the compute power

required at the right time based on its relative priority. In addition, application infrastructure virtualization makes it easier to run applications on different computers, and previously incompatible or legacy applications can be run together on the same physical machine. You will not need to create multiple versions such as Windows or Linux.

Big data platforms designed to support highly distributed, data-intensive applications will run better and faster in a virtual environment. This does not mean that you will want to virtualize all big data–related applications. For example, a text analytics application may run best in a self-contained environment and virtualization would not add any benefit.

Network virtualization

Network virtualization — software-defined networking — provides an efficient way to use networking as a pool of connection resources. Networks are virtualized in a similar fashion to other physical technologies. Instead of relying on the physical network for managing traffic between connections, you can create multiple virtual networks all utilizing the same physical implementation. This can be useful if you need to define a network for data gathering with a certain set of performance characteristics and capacity and another network for applications with different performance and capacity. Limitations in the network layer can lead to bottlenecks that lead to unacceptable latencies in big data environments. Virtualizing the network helps reduce these bottlenecks and improve the capability to manage the large distributed data required for big data analysis.

Processor and memory virtualization

Processor virtualization helps to optimize the processor and maximize performance. Memory virtualization decouples memory from the servers.

In big data analysis, you may have repeated queries of large data sets and the creation of advanced analytic algorithms, all designed to look for patterns and trends that are not yet understood. These advanced analytics can require lots of processing power (CPU) and memory (RAM). For some of these computations, it can take a long time without sufficient CPU and memory resources. Processor and memory virtualization can help speed the processing and get your analysis results sooner.

Data and storage virtualization

Data virtualization can be used to create a platform for dynamic linked data services. This allows data to be easily searched and linked through a unified reference source. As a result, data virtualization provides an abstract service that delivers data in a consistent form regardless of the underlying physical database. In addition, data virtualization exposes cached data to all applications to improve performance.

Storage virtualization combines physical storage resources so that they are more effectively shared. This reduces the cost of storage and makes it easier to manage data stores required for big data analysis.

Data and storage virtualization play a significant role in making it easier and less costly to store, retrieve, and analyze the large volumes of fast and varying types of data. Remember that some big data may be unstructured and not easily stored using traditional methods. Storage virtualization makes it easier to store large and unstructured data types. In a big data environment, it is advantageous to have access to a variety of operational data stores on demand. For example, you may only need access to a columnar database infrequently. With virtualization, the database can be stored as a virtual image and invoked whenever it is needed without consuming valuable data center resources or capacity.

Management and security challenges with virtualization

Virtualized environments need to be adequately managed and governed to realize cost savings and efficiency benefits. If you rely on big data services to solve your analytics challenges, you need to be assured that the virtual environment is as well managed and secure as the physical environment. Some of the benefits of virtualization, including ease of provisioning, can easily lead to management and security problems without proper oversight. Virtualization makes it easy for developers to create a virtual image, or a copy, of a resource. As a result, many companies have implemented virtualization only to find that the number of virtual images spirals out of control. Problems to watch out for include the following:

- Too many virtual images are created, leading to a sharp drop in server and memory performance.
- Lack of control over the life cycle of virtual images leads to the introduction of security vulnerabilities.
- An overabundance of virtual images increases storage costs and reduces cost savings.
- Administrators may increase security risks through either malicious or uninformed management of virtual images.
- Compliance requirements may be compromised if you are not able to accurately monitor virtual infrastructure logs.

Managing Virtualization with the Hypervisor

In an ideal world, you don't want to worry about the underlying operating system and the physical hardware. A *hypervisor* is the technology responsible for ensuring that resource sharing takes place in an orderly and repeatable way. It is the traffic cop that allows multiple operating systems to share a single host. It creates and runs virtual machines. The hypervisor sits at the lowest levels of the hardware environment and uses a thin layer of code (often called a *fabric*) to enable dynamic resource sharing. The hypervisor makes it seem like each operating system has the physical resources all to itself.

In the world of big data, you may need to support many different operating environments. The hypervisor becomes an ideal delivery mechanism for the technology components of the big data stack. The hypervisor lets you show the same application on lots of systems without having to physically copy that application onto each system. As an added benefit, because of the hypervisor architecture, it can load any (or many) different operating systems as though they were just another application. So, the hypervisor is a very practical way of getting things virtualized quickly and efficiently.

You need to understand the nature of the hypervisor. It's designed like a server OS rather than like the Windows OS. Each virtual machine running on a physical machine is called a guest machine. The hypervisor, therefore, schedules the access that guest operating systems have to everything, including the CPU, memory, disk I/O, and other I/O mechanisms. The guest operating systems are the operating systems running on the virtual machines. With virtualization technology, you can set up the hypervisor to split the physical computer's resources. Resources can be split 50/50 or 80/20 between two guest operating systems, for example.

The beauty of this arrangement is that the hypervisor does all the heavy lifting. The guest operating system doesn't care (or have any idea) that it's running in a virtual partition; it thinks it has a computer all to itself.

You find basically two types of hypervisors:

- **Type 1 hypervisors** run directly on the hardware platform. They achieve higher efficiency because they're running directly on the platform.
- **Type 2 hypervisors** run on the host operating system. They are often used when a need exists to support a broad range of I/O devices.

Abstraction and Virtualization

For IT resources and services to be virtualized, they are separated from the underlying physical delivery environment. The technical term for this act of separation is called *abstraction.* Abstraction is a key concept in big data. MapReduce and Hadoop are distributed computing environments where everything is abstracted. The detail is abstracted out so that the developer or analyst does not need to be concerned with where the data elements are actually located.

Abstraction minimizes the complexity of something by hiding the details and providing only the relevant information. For example, if you were going to pick up someone whom you've never met before, he might tell you the location to meet him, how tall he is, his hair color, and what he will be wearing. He doesn't need to tell you where he was born, how much money he has in the bank, his birth date, and so on. That's the idea with abstraction — it's about providing a high-level specification rather than going into lots of detail about how something works. In the cloud, for instance, in an Infrastructure as a Service (IaaS) delivery model, the details of the physical and virtual infrastructure are abstracted from the user.

Implementing Virtualization to Work with Big Data

Virtualization helps makes your IT environment smart enough to handle big data analysis. By optimizing all elements of your infrastructure, including hardware, software, and storage, you gain the efficiency needed to process and manage large volumes of structured and unstructured data. With big data, you need to access, manage, and analyze structured and unstructured data in a distributed environment.

Big data assumes distribution. In practice, any kind of MapReduce will work better in a virtualized environment. You need the capability to move workloads around based on requirements for compute power and storage.

Virtualization will enable you to tackle larger problems that have not yet been scoped. You may not know in advance how quickly you will need to scale.

Virtualization will enable you to support a variety of operational big data stores. For example, a graph database can be spun up as an image.

The most direct benefit from virtualization is to ensure that MapReduce engines work better. Virtualization will result in better scale and performance for MapReduce. Each one of the Map and Reduce tasks needs to be executed independently. If the MapReduce engine is parallelized and configured to run in a virtual environment, you can reduce management overhead and allow for expansions and contractions in the task workloads. MapReduce itself is inherently parallel and distributed. By encapsulating the MapReduce engine in a virtual container, you can run what you need whenever you need it. With virtualization, you increase your utilization of the assets you have already paid for by turning them into generic pools of resources.