IN THIS CHAPTER

- Deconstructing cloud concepts
- Discovering resource pools/cloud models and services
- Evaluating the role of the data center
- Finding out how the public cloud fits and when the private cloud shines

Chapter 1
Understanding the Cloud

There is no debate about it: The cloud is helping to change every business in every industry. The key attributes of the cloud — on-demand access to compute and storage, limitless scaling, and flexible pricing — have allowed startups to challenge well-established industry stalwarts. Likewise, the cloud has enabled established businesses to add new capabilities and transform business process at the speed of change.

The first phase of cloud adoption was mostly about cutting costs and changing technology spend from a capital expense to an operating expense. For many years, cloud vendors were focused on reducing costs for commodity compute and storage. While reducing costs is still a major driver of cloud adoption, companies are now leveraging cloud services to transform their businesses. Most companies have pragmatically adopted a hybrid cloud strategy. They're managing to use multiple public and private cloud services depending on the business requirements. IT leaders are maintaining key mission critical applications in their data centers while taking advantage of innovative Software as a Service (SaaS) applications and cloud infrastructure services. Increasingly, vendors are specializing so that they can provide differentiated value to customers. Some cloud providers offer machine learning services while others are providing governance cloud services. Other vendors are providing a variety of cloud storage services.

In this chapter, we provide an overview of cloud computing, including the basics you need to understand in order to move forward in the world of hybrid and multicloud environments.

Looking at the Ecosystem of Cloud Computing

It's important to understand that an ecosystem of participants define the market. This ecosystem consists of three categories of players:

- **Consumers of services:** These are the end-users that use cloud services in their day-to-day business activities. They may have little understanding of where the service resides or how it is designed; they simply need the capabilities to get the job done.
- **Provider of services:** These cloud providers offer a variety of functions ranging from infrastructure services to applications and tools.
- **Designer of services:** These companies build applications and tools. Often services are intended to work within a specific cloud ecosystem or can augment a packaged cloud application.

Understanding Cloud Concepts

Cloud computing is a method of providing shared computing resources, including applications, computing, storage, networking, development, and deployment platforms as well as business processes. Cloud computing makes computing resources easier to use by providing standardization and automation.

Standardization is the implementation of services using a consistent approach supported by a set of consistent interfaces. Likewise, the cloud generally requires that processes be implemented through the use of automation.

Automation is a process that's triggered based on business rules, resource availability, and security demands. Automation is required to support a self-service provisioning model. To promote efficiency, automation can ensure that after a provisioned service is no longer needed, it is returned to the resource pool. This type of rules-based automation can help with capacity planning and overall workload management.

Most businesses today are already using some kind of cloud service — even if they don't think of it as a cloud. For example, any company that uses Microsoft 365, Slack, or DocuSign service is using a cloud-based service. A company may use online data backup or collaboration services in a commercial cloud. Your organization may also place advertisements or recruit new employees on an open

community cloud like LinkedIn. If your company uses Google's Gmail service, it is using a cloud email service. Many companies are discovering that having Customer Relationship Management (CRM) available as a service is a better way to support the sales team than the traditional on-premises software options.

You should be getting the idea that cloud computing means that everything — from compute power to computing infrastructure and from applications and business processes to personal collaboration — can be delivered to you as a service. To be operational in the real world, the cloud must be implemented with common standardized processes and automation.

REMEMBER

Clouds come in different versions, depending on your needs. There are two primary deployment models of cloud: public and private. Most organizations will use a combination of private computing resources (data centers and private clouds) and public services, where some of the services existing in these environments interact with each other — which is what we call a *hybrid cloud environment*. In addition, many organizations use a variety of public cloud services to support different developer and business units – called a *multicloud environment*. Multicloud has grown in popularity because developers want access to the platform of their choice, and businesses want the flexibility to move between vendors.

The public cloud

The *public cloud* is a set of hardware, networking, storage, services, applications, and interfaces owned and operated by a third party for use by other companies or individuals. These commercial providers create a highly scalable data center that hides the details of the underlying infrastructure from the consumer. Public clouds are viable because they offer many options for computing, storage, and a rich set of other services. With many resources always available, public cloud consumers can quickly select, optimize, and use those resources that match the needs of the applications they will run in the public cloud. Most public cloud providers offer a wide variety of APIs and services, such as security, specialized infrastructure to support specific workloads like Graphic Processer Units (GPUs) for data science, application development pipelines, and other technologies to support customer needs. All of these cloud services are available in an on-demand manner.

TIP

Public cloud vendors are increasingly offering dedicated, non-multi-tenancy instances within their data center. In these instances, you are assigned your own machines and storage within the cloud vendor's data center. Although the workloads are physically isolated, at some point, you will share some networking with other cloud customers. Companies that take this dedicated instance approach typically have governance, compliance, or corporate rules that don't allow multi-tenancy. However, it is important to consider the fact that you will not have the same cost savings of a multi-tenancy approach.

The private cloud

A *private cloud* is a set of hardware, networking, storage, services, applications, and interfaces owned and operated by an organization for the use of its employees, partners, or customers. A private cloud can be created and managed by a third party for the exclusive use of one enterprise. The private cloud is a highly controlled environment not open for public consumption. Thus, a private cloud sits behind a firewall. The private cloud is highly automated with a focus on governance, security, and compliance. Automation replaces more manual processes of managing IT services to support customers. In this way, business rules and processes can be implemented inside software so that the environment becomes more predictable and manageable.

Increasingly, public cloud vendors are packaging their cloud services into appliances that can be installed within a customer's on-premises data center behind the firewall. The appliance typically contains access to all of the cloud services that the cloud vendor offers on the public cloud. The consumption models for these appliance based on premises public clouds can vary — the vendor may manage and own the appliance and bill the client in the same way that they bill public cloud use, or the customer may own and maintain the appliance. This model of public cloud capabilities behind the firewall is quickly gaining traction. In these cases, businesses get the scalability, ease of use, cost model and familiarity of a cloud environment while keeping data and workloads on premises.

The hybrid and multicloud model

A *hybrid cloud* is a combination of a private cloud combined with the use of public cloud services where the two cloud environments work together to solve business problems. The goal is to create a hybrid cloud environment that can combine services and data from a variety of cloud models to create a unified, automated, and well-managed computing environment. In a well-orchestrated hybrid cloud environment, end-users won't think about whether their using on premises or cloud services — it will all just be a technology service.

In addition to the hybrid cloud, *multicloud* is when two or more public cloud are being used within an organization. Many businesses initially found that they had a multicloud environment because different development teams or business units were choosing to use varying public clouds. As you can imagine, as businesses found themselves using multiple clouds, finance, operations, and IT teams needed a way to gain visibility, control, and choice between clouds. Therefore, multicloud management is emerging as an important consideration.

Combining multiple public services with private clouds and the data center is the definition of *corporate computing.* Not all companies that use some public and some private cloud services have a hybrid or multi cloud. Rather, a hybrid or multicloud environment is when multiple public and/or private services are used together to create value. In the following circumstances, a computing environment is not a hybrid or multicloud:

- If a few developers in a company use a public cloud service to prototype a new application that is completely disconnected from the private cloud or the data center, the company does not have a hybrid environment.
- If a company is using a SaaS application for a project but there is no movement of data from that application into the company's data center, the environment is not hybrid.
- If different divisions within an enterprise are standardized on different public cloud infrastructures, but each division only utilizes and manages a single public cloud.

A cloud is hybrid or multi in the following situations:

- If a company uses a public development platform that sends data to a private cloud or a data center–based application, the cloud is hybrid.
- When a company leverages a number of SaaS applications and moves data between private or data center resources, the cloud is hybrid.
- When a business process is designed as a service so that it can connect with environments as though they were a single environment, the cloud is hybrid.
- When a SaaS analytics platform is used and data from multiple clouds sources is ingested.
- When your organization can move workloads to different public clouds based on cost or performance concerns.

Cloud Computing Elements: Resource Pools/Cloud Models and Services

Now that you have a context for the types of cloud environments, it's important to understand the common elements required to make clouds functional. In this section, we give you the basics of what you need to know. Figure 1–1 illustrates the related elements that come together to create clouds. On the bottom of the

diagram is a set of *resource pools* that feed a set of cloud delivery services. On the top of the diagram are the common service elements needed to support these delivery models.

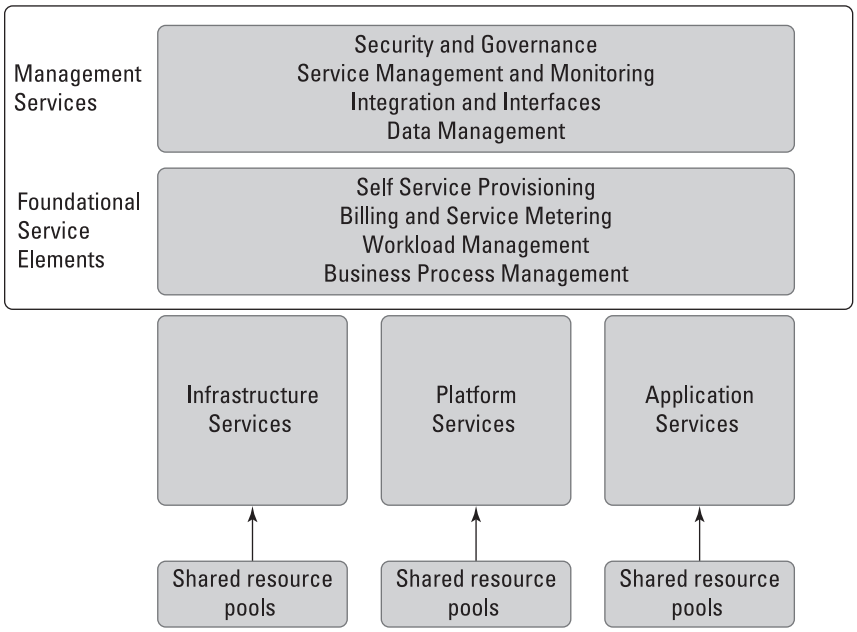

FIGURE 1-1: How different cloud elements fit together.

So far we have been talking about resources that represent infrastructure technology like processors, storage, and networking. But resources also include software services like databases, identity management, and email servers. Those software services are also available in cloud environments, and they are typically constructed as a shared, *multi-tenant* service. Multi-tenancy is a computing architecture that allows customers to share computing resources in the cloud. Although customers are sharing resources, their individual configurations and data are isolated.

For example, say that you're a cloud provider. If each customer had their own dedicated server, storage device, and networking equipment, scaling your cloud business to support thousands of customers would be hugely expensive. In addition, you'd have to offer your cloud services at a very high cost. On the other hand, if the cloud vendor uses a multi-tenancy architecture, they can have multiple customers using shared infrastructure. Similarly, cloud software vendors use a multi-tenancy architecture so that they don't have to replicate their software for every single user. Your software data, settings, and preferences are all saved, but the underlying base software is shared in a multi-tenancy approach.

To make resource pooling work, each pooled element needs to be written with service-oriented constructs in mind. Consequently, each resource is written as an independent service without dependencies and with well-defined interfaces. For more details on cloud services, turn to Chapter 6.

Cloud delivery models

Understanding the foundations of cloud computing calls for an understanding of cloud delivery models. In this section, we focus on the models that represent computing environments:

- Infrastructure as a Service
- Platform as a Service
- Software as a Service

Part 2 of this book has a chapter on each of these models, but here we cover the basics of the different cloud computing delivery models. These are illustrated as infrastructure services, platform services, and software services in Figure 1-1, earlier in this chapter.

Infrastructure as a Service

Infrastructure as a Service (IaaS) is the delivery of services, including an operating system, storage, networking, and various utility software elements, on a request basis. The easiest way to think of IaaS is that it provides a virtual server that is equivalent to a physical server — you have to select an operating system (for example, Linux, Windows, and so on), and everything "up the stack" to the applications that will run.

IaaS has both public and private versions. In the public IaaS, the public cloud provider creates the infrastructure and resources that consumer can use. The user simply needs a credit card to acquire these resources. When that user stops paying, the resource may disappear. In a private IaaS service, it is usually the IT organization or an integrator who creates an infrastructure and resources that internal users and sometimes business partners can use on demand. Whereas criteria for a public cloud are based primarily on the ability to pay for a service, a private service applies company policy to a service request. Some customers will bring their own tools and software to create applications. For more details on IaaS, turn to Chapter 8.

Platform as a Service

Platform as a Service (PaaS) is a mechanism for combining IaaS with an abstracted set of middleware services, software development, and deployment tools that

allow the organization to have a consistent way to create and deploy applications on a cloud or on-premises environment. The easiest way to think about PaaS is that it's an IaaS, but the operating system and development tools are already in place. Because a PaaS environment is ready for development, productivity and time to value is greatly increased.


TIP

Many PaaS environments are anchored to software platforms. For example, Salesforce is a generalized Customer Relationship Management (CRM) to help companies manage their relationships and interactions with customers and prospective customers. Salesforce has a large PaaS platform so that partners with specific domain expertise can use the Salesforce PaaS platform to build industry specific CRMs specific industries, including pharmaceutical sales, education, retail clothing, and food sales.

A PaaS offers a consistent set of programming and middleware services that ensure developers have a well-tested and well-integrated way to create applications in a cloud environment. A PaaS environment brings development and deployment together to create a more manageable way to build and deploy applications. A PaaS requires an Infrastructure service. For more on PaaS, see Chapter 10.

Software as a Service

Software as a Service (SaaS) is a business application created and hosted by a provider in a multi-tenant model. Some of the most popular SaaS applications include Google's G Suite Apps, ServiceNow, Salesforce, Adobe Create Cloud, and Zendesk. Customers typically pay for the SaaS service per user on a monthly or yearly contract model. The SaaS application sits on top of both a Platform as a Service and foundational Infrastructure services. However, customers do not typically care about the underlying infrastructure and platform services. Instead, customers are concerned about the functionality, performance, and availability, and security of the applications. To read more details on SaaS, turn to Chapter 9.

The computing resources life cycle

A fundamental notion of the cloud, both public and private, is that consumers utilize computing resources only when they want to and are only charged for the resources they actually use, for the time they use those resources. In that sense, they rent resources for time periods they desire. Contrast that with the physical data center where computers, storage, and other resources are purchased and then used for the life of those resources.

To meet consumer needs, cloud service providers need to design their computing platforms to respond immediately to whatever consumers request. Since the providers don't know when consumers will make their requests, how many resources

they will need, the size of those resources, and how long they will keep those resources, the design must keep many resources available for use at all times. Moreover, the cloud provider must track all resource usage down to fractions of a second so that consumers can be billed for exactly what they used on a pay-as-you-go basis. When consumers finish using resources, the cloud platform must return those resources to the set of available resources so that other consumers can use them.

In some situations, a service provider can't anticipate the needs of a customer. Therefore, it is common for a service provider to add capacity from a third-party service provider. Typically, consumers are unaware that they are dealing with an additional cloud service provider.

Understanding Self-Service Provisioning and Elasticity

Self-service provisioning is one of the most important capabilities of cloud computing. With self-service, cloud consumers can use a website in the cloud to select and purchase cloud services, configure them, launch them into the cloud environment, and start using them within minutes or perhaps even seconds. In the traditional data center model, that same consumer might have to file a request with IT operations for equipment or software, go through approval and payment processes, and then wait while IT procures the equipment, installs it, and configures it, and finally turns it over to the requesting consumer for use. The data center procurement process can take days, and often weeks.

For example, consider a developer with a new application that requires testing with a new server. The delay in waiting for the server not only holds up the development process, but delays the time to market for the application, costing the business for lost revenue and sometimes risking market share by not being a first mover. We go into more detail about how cloud computing helps with software development and operations in Chapter 11.

Closely related to self-service is the concept of *elasticity*, in which cloud resources can automatically change their own provisioned size. The basic meaning of elasticity is that many cloud resources can be selected at a specific size or quantity, but when usage of those resources starts to approach the original size, they automatically increase their own size.

For example, a cloud application that will store customer pictures can select an elastic file system to store the pictures. The application can specify a starting size

of the filing system, but as more pictures are added and the total storage exceeds the originally configured size, the file system will automatically add more storage so there is again room for storing files. In this sense, the storage resource is elastic, growing to accommodate more files and, if files are removed, shrinking the storage as well. Because cloud consumers pay only for the resources they actually use, elasticity works to the consumer's financial, as well as technical, benefit by reducing costs.

Establishing a Dynamic Life Cycle across Workloads and Data

A cloud isn't a single unified environment; rather, it's a combination of resources that may be spread across systems and geographies. The cloud is a federated environment that brings together resources so that they can work together. To make this happen in an organized manner requires an organization of workloads. A *workload* is an independent service or collection of code that can be executed. So, you need to think about the cloud as a group of workloads that are managed as though they were a single cohesive environment.

It is important in a cloud environment that workloads be designed to support the right task with the right cloud services. For example, some workloads will need to be placed in a private cloud because they require fast transaction management with legacy systems that have not been moved to the cloud. Other workloads may not have legacy constraints and can be placed in a public cloud.

When organizations begin thinking about cloud computing as a strategy, they must do more than simply go to a public or private resource and leverage those services. Like any computing environment, cloud computing requires that workloads be balanced and managed. Today's reality includes many cloud providers and environments, and a strategy should include what is called multicloud — the strategic use of multiple cloud providers (we cover this topic in detail in Chapter 12). Within a well-designed cloud or multicloud environment, workloads and data can move across multiple cloud providers, geographic regions, or service levels. Because, in the real world, you will use a combination of services, it's important to think not just about an individual workload but also about a combination of workloads and how they interact with each other and with collaborators.

Other issues to consider include locating workloads and data in the same region as your primary customers to optimize performance, understanding the life cycle of your application and data to select optimizations based in increasing — or decreasing — demand, and even moving workloads to alternate cloud providers when the differential of costs, reliability, or features make it advantageous.

Management Services

Management services are mandatory for ensuring that the operation of your cloud workloads and resources meets constituent needs (whether they are customer, employee or partner) needs. This is the case regardless of the cloud deployment and delivery model. Some core services are illustrated at the top of Figure 1-1, which appears earlier in this chapter.

Network monitoring and management is critical because outages and slowdowns can have dramatic impact on the customer experience of cloud applications. Application and workload health monitoring can provide warnings of impending problems and can inform a support organization so they can provide the best service to customers. Security and governance are key services that ensure your applications and data are protected. Data management in a hybrid environment is also critical since data will be moving among cloud and physical environments.

Finally, since various services will need to be integrated in a hybrid or multicloud model, the interfaces between clouds are also important.

The Changing Role of the Data Center

What happens to the data center when companies begin to implement hybrid clouds? First, the data center does not go away. After all, almost all medium-size and large companies run their own data center — which is how many companies operate their systems of record, including accounting systems, inventory records, and line of business applications, to name a few. Many data centers have grown in an unplanned manner over many decades. The typical data center supports different hardware architectures, operating systems, applications, and hundreds, if not thousands, of different tools. To make matters worse, a lot of the money spent in supporting a data center is used for maintenance of existing systems, heating, air conditioning, floor space, and labor.

So, it's not surprising that many companies have taken the time to streamline their data centers through technologies, such as server virtualization. In essence, virtualization decouples the software from the hardware. In *decoupling*, the software is put into a virtual machine (much like in the cloud) so it's isolated from the underlying operating system. With the use of virtualization, data center management can more easily and efficiently manage the way applications are placed on servers.

However, even though IT has made the data center more efficient, cloud computing has made it apparent that more can be done to transform computing. Organizations are taking a hard look at what the centralized data center is well suited for

in the light of what they have learned about cloud computing. A comprehensive strategy for managing workloads and data will consider the traditional data center as well as private, public, hybrid, and multicloud environments.

TIP

It wasn't long ago that the cloud was believed to be less secure than physical data centers. However, most commercial public clouds provide sophisticated levels of security. In a traditional data center, many attacks come from internal bad actors or employees who are well-meaning, but make mistakes that leave your company open to an intrusion.

REMEMBER

There are still good reasons to run certain workloads on premises — for example, legacy applications that are not prepared to take advantage of the scalability and elasticity of the cloud, or applications that have dependencies on external services or hardware that are not available in the cloud. In many cases, there many simply not be a reasonable business case to move an application to the cloud. Some data is also better kept in a physical data center, especially data which is subject to regulatory governance and compliance. Also, the growing issue of *data sovereignty* — the concept that information which is stored in digital form is subject to the laws of the country in which it is located — is a driver for keeping data in a physical data center.

Evolution of the data center into a private cloud

Just as businesses saw the value of virtualization in streamlining IT operations in the data center, the cloud is seen as the next step along that path.

IT organizations have discovered that it's much more efficient and effective to create private cloud services for developers to create new applications and services. Therefore, companies are setting up a highly automated computing environment enabled with a self-service portal. This portal is often designed with business process rules that dictate what services a developer or an authorized partner can use. For example, a developer beginning to develop a new application may be permitted to use the Java language with the Eclipse IDE, specific types of middleware, and a specified amount of computing capacity and storage. Once the project is completed, there may be a rule that automatically de-provisions the resources. The private cloud service is intended to support an organization's need for speed and agility based on fast-changing company requirements. Having a private cloud available for projects allows a company to easily experiment with new ideas and new applications without having to request funding for a project that might not become a reality.

Seeing how the public cloud fits

Some business initiatives are good matches for the public cloud. In the next section, we discuss situations that are more appropriate for a private cloud.

First, say that your company is a retailer with a vibrant and well-used SaaS portal for selling products online. The biggest event for the company is its yearly 50 percent–off sale. Although their public cloud has the elastic resources to scale up the portal to handle the increased demand during the sale, this year they plan to expand their sale to Europe and the Far East. Each of those regions has its own inventory to fulfill orders locally, but sales and inventory management is handled by the SaaS portal. Your company selected its cloud provider partly because they run worldwide cloud environments. The cloud provider makes it easy to replicate applications and data to different regions, so it was simple to set up the SaaS application in Europe and the Far East, ensuring that performance for customers in those regions is as good as in their home market. Furthermore, it was easy to take the SaaS application down in those other regions after the sale. Because your company can rely on its public cloud provider, it doesn't have to buy and install additional servers for the sale or have to remove them afterward, thus saving time and money.

Your company has selected two different software applications from service providers to replace existing on-premises applications. One application is a customer relationship management (CRM) that allows the sales force to easily get access to prospective and current customer information from a cloud-based public service. In addition, your company uses a human resources management platform as a service. The company has implemented integration software that allows data to be managed among their on-premises data center and the two cloud SaaS environments, forming a hybrid cloud. Both of these SaaS applications have enabled your company to avoid purchasing additional hardware and software that would require IT management. In addition, because the sales team can access their data much faster from any device they're using, from a business perspective, the sales team's performance is much more effective.

Your company is in a new market where getting a series of new services operational quickly has the potential to leapfrog the competition. Your company hasn't been in business very long. Because the company can use a public service, it can make services available before more established companies can act. At the same time, the company can use its on-premises data center in a hybrid cloud configuration to monitor the effectiveness of services, manage data privately, and combine with other services that aren't visible to customers. All the information collected in the on-premises data center is processed to create analytics that your company sees an opportunity to market and make available in the public cloud, further enhancing its revenue.

Knowing when the private cloud shines

In some situations, the private cloud is a better choice than a public cloud, for both applications and data.

Say that your financial services company's products and services are offered in a portal to your business partners. These products are key to a company's revenue. For example, your company might offer a key business service that is purchased by various banks around the globe. Because of the worldwide reach of your service, it makes sense for the company to establish its portal in a public cloud. However, your business service includes sensitive information that must, by some government's legal requirements, be stored in a physical data center. An architected private cloud that maintains the sensitive financial data is ideal for the financial services company. The resulting hybrid cloud connects the public cloud's portal with the private cloud financial data. The private cloud provides the scalability required for the quantity of data being managed, as well as meeting the legal obligations to store the regulated data in the country where the business is conducted.

A decade ago, your company's IT department realized that their physical data center was over-provisioning their hardware to meet growing computing demands, and decided to migrate their data center applications and data to a virtualized environment. At the time, it was a successful initiative, and they reduced costs by cutting back on over-provisioning and increasing their ability to manage their workloads and data. However, over time, the ease of creating new virtual machines (for example, when development would create new test and staging environments) led to "virtual machine sprawl" where management became impossible. Although their corporate goal became to move their applications to the public cloud, they realized that not only did they need to get control of their VM sprawl, some applications and systems were not cloud-ready. They decided to create a private cloud as a stepping stone to the public cloud. The private cloud allowed them to upgrade their applications to be cloud-native before putting them in the public cloud. It also provided a safe environment for getting control of their virtual machines. Using the private cloud as a testing and staging environment for the move to the public cloud minimized risks of failures in the public cloud and streamlined their management approaches. A further benefit was that after their applications were moved from the private to public cloud, they were left with a modern private cloud in their data center that already had the beginnings of a hybrid cloud configuration from their migration to the public cloud.