Chapter 1

Grasping the Fundamentals

In This Chapter

- Doing your computing on the cloud
- Seeing what the cloud's made of
- Comparing the cloud to tradition
- Driving your business

In a dynamic economic environment, your company's survival may depend on your ability to focus on core business and adapt quickly. Yesterday's profitable business model can't be counted on to translate into future growth and profits. As your business adapts to changing government and industry regulations, evaluates new business partnerships, and anticipates competitive threats, IT needs to help the business find new ways to respond.

At the same time, plans for change must often be made in the context of limited resources for finances, people, technology, and power. In this chapter, we introduce you to cloud computing — what it is and how it helps companies rethink how they deploy technology.

While there are a lot of technical considerations, keep in mind the fundamental truth: Cloud computing is a business and economic model. Is cloud computing a replacement for the traditional data center? The answer is complicated. In some cases, yes; in some cases, no.

Are we suggesting that the traditional data center goes away to be replaced with a cloud? Not necessarily. Sometimes the traditional data center is the best fit. However, for business agility and economic reasons, the cloud is becoming an increasingly important option for companies. We see cloud computing as the foundation for the industrialization of computing. Yes, it is that important.

Considering Perspectives

In this book, we look at cloud computing from three perspectives: the strategy from both the customer and the provider's point of view, business and economic considerations, and the technical underpinnings. We also examine how companies are using the cloud to control IT expenditures as they prepare to move to a service-centric world.

Many players make up the world of cloud computing:

- The **vendors** providing applications and enabling technology, infrastructure, hardware, and integration
- The **partners** of these vendors that are creating cloud services offerings and providing support services to customers
- The **business leaders** themselves who are either using or evaluating various types of cloud computing offerings

This book addresses each of these audiences because they're all a fundamental part of this fabric of the future of computing.

Computing on the Cloud

What is cloud computing? Cloud computing is the next stage in evolution of the Internet. The *cloud* in cloud computing provides the means through which everything — from computing power to computing infrastructure, applications, business processes to personal collaboration — can be delivered to you as a service wherever and whenever you need.

Cloud computing is offered in different forms:

- Public clouds
- Private clouds
- Hybrid clouds, which combine both public and private

In general the cloud — similar to its namesake of the cumulus type — is fluid and can easily expand and contract. This *elasticity* means that users can request additional resources on demand and just as easily *deprovision* (or release) those resources when they're no longer needed. This elasticity is one of the main reasons individual, business, and IT users are moving to the cloud.

In the traditional data center it has always been possible to add and release resources. However, this process couldn't be done in an automated or self-service manner.

This evolution to cloud computing — already underway — can completely change the way companies use technology to service customers, partners, and suppliers. Some businesses already have IT resources almost entirely in the cloud. They feel that the cloud model provides a more efficient, cost-effective IT service delivery.

This doesn't mean that all applications, services, and processes will necessarily be moved to the cloud. Many businesses are much more cautious and are taking a hard look at their most strategic business processes and intellectual property to determine which computing assets need to remain under internal company control and which computing assets could be moved to the cloud.

Defining the Cloud

The *cloud* itself is a set of hardware, networks, storage, services, and interfaces that enable the delivery of computing as a service. *Cloud services* include the delivery of software, infrastructure, and storage over the Internet (either as separate components or a complete platform) based on user demand.

The world of the cloud has lots of participants:

- The **end user** doesn't really have to know anything about the underlying technology. In small businesses, for example, the cloud provider becomes the de facto data center. In larger organizations, the IT organization oversees the inner workings of both internal resources and external cloud resources.
- **Business management** needs to take responsibility for overall governance of data or services living in a cloud. Cloud service providers must provide a predictable and guaranteed service level and security to all their constituents.
- The **cloud service provider** is responsible for IT assets and maintenance.

Therefore, we have written this book to include the concerns of all the players in the evolving cloud ecosystem.

Cloud services must enable *multi-tenancy* — different companies sharing the same underlying resources. This topic is discussed further in Chapter 12.

Companies are finding some important new value in cloud services. The cloud can eliminate many of the complex constraints from the traditional computing environment, including space, time, power, and cost.

Cloud services like social networks (such as Facebook or LinkedIn) and collaboration tools (like video conferencing, document management, and webinars) are changing the way people in businesses access, deliver, and

understand information. Cloud computing infrastructures make it easier for companies to treat their computing systems as a pool of resources rather than a set of independent environments that each has to be managed.

Overall, the cloud embodies the following four basic characteristics:

- Elasticity and the ability to scale up and down
- Self-service provisioning and automatic deprovisioning
- Application programming interfaces (APIs)
- Billing and metering of service usage in a pay-as-you-go model

Each of these characteristics is described in more detail in the following sections.

Elasticity and scalability

The service provider can't anticipate how customers will use the service. One customer might use the service three times a year during peak selling seasons, whereas another might use it as a primary development platform for all of its applications.

Therefore, the service needs to be available all the time (7 days a week, 24 hours a day) and it has to be designed to scale upward for high periods of demand and downward for lighter ones. *Scalability* also means that an application can scale when additional users are added and when the application requirements change.

This ability to scale is achieved by providing *elasticity*. Think about the rubber band and its properties. If you're holding together a dozen pens with a rubber band, you probably have to fold it in half. However, if you're trying to keep 100 pens together, you will have to stretch that rubber band. Why can a single rubber band accomplish both tasks? Simply, it is elastic and so is the cloud.

In Chapter 2, we give you some concrete examples of how providers are using this characteristic.

Self-service provisioning

Customers can easily get cloud services without going through a lengthy process. The customer simply requests an amount of computing, storage, software, process, or other resources from the service provider. Chapter 7 explains this process in detail.

Contrast this on-demand response with the process at a typical data center. When a department is about to implement a new application, it has to submit a request to the data center for additional computing hardware, software, services, or process resources. The data center gets similar requests from departments across the company and must sort through all requests and evaluate the availability of existing resources versus the need to purchase new hardware. After new hardware is purchased, the data center staff has to configure the data center for the new application. These internal procurement processes can take a long time, depending on company policies.

Of course, nothing is as simple as it might appear. While the on-demand provisioning capabilities of cloud services eliminates many time delays, an organization still needs to do its homework. These services aren't free; needs and requirements must be determined before capability is automatically provisioned.

Application programming interfaces (APIs)

Cloud services need to have standardized APIs. These interfaces provide the instructions on how two application or data sources can communicate with each other.

A standardized interface lets the customer more easily link a cloud service, such as a customer relationship management system with a financial accounts management system, without having to resort to custom programming. For more information on standards see Chapter 14.

Billing and metering of services

Yes, there is no free lunch. A cloud environment needs a built-in service that bills customers. And, of course, to calculate that bill, usage has to be *metered* (tracked). Even free cloud services (such as Google's Gmail or Zoho's Internet-based office applications) are metered.

In addition to these characteristics, cloud computing must have two overarching requirements to be effective:

- A comprehensive approach to service management
- A well-defined process for security management

Performance monitoring and measuring

A cloud service provider must include a service management environment. A *service management environment* is an integrated approach for managing your physical environments and IT systems. This environment must be able to maintain the required service level for that organization.

In other words, service management has to monitor and optimize the service or sets of services. Service management has to consider key issues, such as performance of the overall system, including security and performance. For example, an organization using an internal or external email cloud service would require 99.999 percent uptime with maximum security. The organization would expect the cloud provider to prove that it has met its obligations.

Many cloud service providers give customers a dashboard — a visualization of key service metrics — so they can monitor the level of service they're getting from their provider. Also, many customers use their own monitoring tools to determine whether their service level requirements are being met.

Security

Many customers must take a leap of faith to trust that the cloud service is safe. Turning over critical data or application infrastructure to a cloud-based service provider requires making sure that the information can't be accidentally accessed by another company (or maliciously accessed by a hacker).

Many companies have compliance requirements for securing both internal and external information. Without the right level of security, you might not be able to use a provider's offerings. For more details on security, see Chapter 15.

Comparing Cloud Providers with Traditional IT Service Providers

Traditional IT service providers operate the hardware, software, networks, and storage for its clients. While the customer pays the licensing fees for the software, the IT service provider manages the overall environment. The service provider operates the infrastructure in its own facilities. With the traditional IT service provider, the customer signs a long-term contract that specifies mutually agreed-upon service levels. These IT providers typically customize an environment to meet the needs of one customer.

In the cloud model, the service provider might still operate the infrastructure in its own facilities (except in the case of a private cloud, which we discuss in Chapter 9). However, the infrastructure might be *virtualized* across the globe, meaning that you may not know where your computing resources, applications, or even data actually reside. (We talk more about virtualization in Chapter 17.) Additionally, these service providers are designing their infrastructure for scale, meaning that there isn't necessarily a lot of customization going on. (We talk more about the scale issue in Chapter 13.)

Addressing Problems

There is an inherent conflict between what the business requires and what data center management can reasonably provide. Business management wants optimal performance, flawless implementation, and 100 percent uptime. The business leadership wants new capability to be available immediately, frequent changes to applications, and more accessibility to quality data in real time — but their organizations have limited budgets.

Getting on board with cloud computing

Although opinions differ about how quickly technology will migrate to the cloud, without doubt the interest level is high. Lots of business folks are asking questions about the cloud approach when they hear about the data center efficiencies achieved by companies like Amazon (`www.amazon.com`) and Google (`www.google.com`).

For example, a smart CEO was under a lot of pressure to improve profitability by cutting capital expenditures. One day he read an article about the economic advantages of cloud computing in a business journal and began to wonder, "Hey, if Amazon can offer computing on demand, why can't our own IT department act like that?" The CEO paid a visit to the CIO and asked that very question. The CIO wasn't quite sure how to answer his boss. His only reply was that things are more complicated than that. The CIO pointed out issues related to data security and privacy. In addition, there are applications running in the data center that are one-of-a-kind and not easily handled. At the same time, he recognized that the department needed to provide better service to internal customers. The CIO did agree that there were other areas of IT that might be appropriate for the cloud model. For example, areas such as testing, software development, storage, and email were good candidates for cloud computing.

Over time, it became easier for IT to add hardware to the data center rather than to focus on making the data center itself more effective. And this plan worked. By pouring more resources into the data center, IT ensured that critical applications wouldn't run out of resources. At the same time, these companies built or bought software to meet business needs. The applications that were built internally were often large and complex. They had been modified repeatedly to satisfy changes without regard to their underlying architecture.

Between managing a vast array of expanding hardware resources combined with managing huge and unwieldy business software, IT management found itself under extraordinary pressure to become much more effective and efficient.

This tug of war between the needs of the business and the data center constraints has caused friction over the past few decades. Clearly, need and money must be balanced. To meet these challenges, there have been significant technology advancements including virtualization (see Chapter 17), service-oriented architecture (see Chapter 19), and service management (see Chapter 20). Each of these areas is intended to provide more modularity, flexibility, and better performance for IT.

While these technology enablers have helped companies to become more efficient and cost effective, it isn't enough. Companies are still plagued with massive inefficiencies. The promise of the cloud is to enable companies to improve their ability to leverage what they've bought and make use of external resources designed to be used on demand.

We don't want to give you the idea that everything will be perfect when you get yourself a cloud. The world, unfortunately, is more complicated than that. For example, complex, brittle applications won't all be successful if they are just thrown up on the cloud. Virtualization adds performance implications. And many of these applications lack an architecture to achieve scale. A database-bound application will remain database bound, regardless of the additional compute resources beneath it.

Discovering the Business Drivers for Consuming Cloud Services

In the beginning of this chapter, we name reasons companies are thinking about cloud services and some of the pressures coming from management. Clearly, business management is under a lot of pressure to reduce costs while providing a sophisticated level of service to internal and external customers. In this section, we talk about the benefits of cloud services.

Supporting business agility

One of the most immediate benefits of cloud-based infrastructure services is the ability to add new infrastructure capacity quickly and at lower costs. Therefore, cloud services allow the business to gain IT resources in a self-service manager, thus saving time and money. By being able to move more quickly, the business can adapt to changes in the market without complex procurement processes.

A typical cloud service provider has *economies of scale* (cost advantages resulting in the ability to spread fixed costs over more customers) that the typical corporation lacks. As mentioned earlier, the cloud's self-service capability means it's easier for IT to add more *compute cycles* (more CPU resources added on an incremental basis) or storage to meet an immediate or intermittent needs.

With the advent of the cloud, an organization can try out a new application or develop a new application without first investing in hardware, software, and networking.

Reducing capital expenditures

You might want to add a new business application, but lack the money. You might need to increase the amount of storage for various departments. Cloud service providers offer this type of capability at a prorated basis. A cloud service vendor might rent storage on a per-gigabyte basis.

Companies are often challenged to increase the functionality of IT while minimizing capital expenditures. By purchasing just the right amount of IT resources on demand the organization can avoid purchasing unnecessary equipment. There are always trade-offs in any business situation.

A company may significantly reduce expenses by moving to the cloud and then may find that its operating expenses increase more than predicted. In other situations, the company may already have purchased significant IT resources and it may be more economically efficient to use them to create a private cloud. Some companies actually view IT as their primary business and therefore will view IT as a revenue source. These companies will want to invest in their own resources to protect their business value.