Chapter 6

Examining the Cloud and Big Data

In This Chapter

- Defining the cloud
- Examining cloud deployment and delivery models
- Understanding why the cloud is an imperative for big data
- Looking at how the cloud can be used with big data deployments

The power of the cloud is that users can access needed computing and storage resources with little or no IT support or the need to purchase more hardware or software. One of the key characteristics of the cloud is elastic scalability: Users can add or subtract resources in almost real time based on changing requirements. The cloud plays an important role within the big data world. Dramatic changes happen when these infrastructure components are combined with the advances in data management. Horizontally expandable and optimized infrastructure supports the practical implementation of big data.

In this chapter, we review the fundamentals of the cloud in the context of what it means for big data. Then we discuss how and why the cloud is often so ideal for various use cases for big data.

Defining the Cloud in the Context of Big Data

Cloud computing is a method of providing a set of shared computing resources that include applications, computing, storage, networking, development, and deployment platforms, as well as business processes. Cloud computing turns traditional siloed computing assets into shared pools of resources based on an underlying Internet foundation. In cloud computing, everything, from compute power to computing infrastructure and from applications and business processes to data and analytics, 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.

If you want to find out a lot more about the cloud, we recommend that you read another book we have written, *Hybrid Cloud For Dummies* (published by John Wiley & Sons, Inc.).

Many businesses leverage cloud services for everything from backup to Software as a Service (SaaS) options such as customer relationship management (CRM) services. With the growth of mobile computing, more consumers, professionals, and corporations are creating and accessing data with cloud-based services. The average consumer may be sent an online coupon for a favorite store; a quality control manager in a manufacturing plant might collect sensor data from a variety of machines to determine whether a quality problem exists. These scenarios are predicated on the cloud-based data services infrastructure.

A popular example of the benefits of cloud supporting big data can be noted at both Google and Amazon.com. Both companies depend on the capability to manage massive amounts of data to move their businesses forward. These providers needed to come up with infrastructures and technologies that could support applications at a massive scale. Consider Gmail and the millions upon millions of messages that Google processes per day as part of this service. Google has been able to optimize the Linux operating system and its software environment to support e-mail in the most efficient manner; therefore, it can easily support hundreds of millions of users. Even more importantly, Google is able to capture and leverage the massive amount of data about both its mail users and its search engine users to drive the business.

Likewise, Amazon.com, with its IaaS data centers, is optimized to support these workloads so that Amazon can continue to offer new services and support a growing number of customers without breaking the bank. To grow its retail business, Amazon must be able to manage data about its merchandise, its buyers, and its channel of partner merchants. Targeted advertising based on customer buying patterns is critical to the company's success. These companies now offer a range of cloud-based services for big data that we talk about later in this chapter.

Understanding Cloud Deployment and Delivery Models

Two key cloud models are important in the discussion of big data — public clouds and private clouds. For those organizations that adopt cloud deployment and delivery models, most will use a combination of private computing resources (data centers and private clouds) and public services (operated by an external company for the shared use of a variety of customers who pay a per-usage fee). How these companies balance public and private providers depends on a number of issues, including privacy, latency, and purpose. It is important to understand these environments and what they mean for a potential big data deployment. In that way, you can determine whether you might want to use a public cloud IaaS (described later) — for example, for your big data projects — or if you want to continue to keep all your data on premises. Or, you might want to use a combination of both. So, we outline these deployment and delivery models first and then talk more about what they mean to big data.

Cloud deployment models

The two types of deployment models for cloud computing are public and private. These are offered for general purpose computing needs as opposed to specific types of cloud delivery models. We examine the delivery models later in the chapter. In the meantime, take a look at the differences between public and private cloud models and how you might use them.

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 and 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 typically manage relatively repetitive or straightforward workloads. For example, electronic mail is a very simple application. Therefore, a cloud provider can optimize the environment so that it is best suited to support a large number of customers, even if it saves many messages.

Likewise, public cloud providers offering storage or computing services optimize their computing hardware and software to support these specific types of workloads.

In contrast, the typical data center supports so many different applications and workloads that it cannot be easily optimized. A public cloud can be very effective when an organization is executing a complex data analysis project and needs extra computing cycles to handle the task. In addition, companies may choose to store data in a public cloud where the cost per gigabyte is relatively inexpensive when compared to purchased storage. The overriding issues with public clouds for big data are the security requirements and the amount of latency that is acceptable.

All public clouds are not the same. Some public clouds are scalable managed services with a high level of security and a high level of service management.

Other public clouds are less robust and less secure, but they are much less expensive to use. Your choice will depend on the nature of your big data projects and the amount of risk you can assume.

The private cloud

A private cloud is a set of hardware, networking, storage, services, application, and interfaces owned and operated by an organization for the use of its employees, partners, and 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, the 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 service to support customers. In this way, business rules and processes can be implemented inside software so that the environment becomes more predictable and manageable. If organizations are managing a big data project that demands processing massive amounts of data, the private cloud might be the best choice in terms of latency and security.

TIP

A *hybrid* cloud is a combination of a private cloud combined with the use of public cloud services with one or several touch points between the environments. The goal is to create a well-managed cloud environment that can combine services and data from a variety of cloud models to create a unified, automated, and well-managed computing environment.

Cloud delivery models

In addition to the cloud deployment models discussed previously, a number of cloud delivery models also exist. Four of the most popular are described in the following sections.

Infrastructure as a Service

Infrastructure as a Service (IaaS) is one of the most straightforward of the cloud computing services. IaaS is the delivery of computing services including hardware, networking, storage, and data center space based on a rental model. The consumer of the service acquires a resource and is charged for that resource based on amount used and the duration of that usage. You find both public and private versions of IaaS. In the public IaaS, the user utilizes a credit card to acquire these resources. When the user stops paying, the resource disappears. In a private IaaS service, it is usually the IT organization or an integrator who creates the infrastructure designed to provide resources on demand for internal users and sometimes business partners.

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. A PaaS offers a consistent set of programming or middleware services that ensure that 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, deploy, and scale applications. A PaaS requires an IaaS.

Software as a Service

Software as a Service (SaaS) is a business application created and hosted by a provider in a multitenant model. *Multitenancy* refers to the situation where a single instance of an application runs in a cloud environment, but serves multiple client organizations (tenants), keeping all their data separate. Customers pay for the service per user either on a monthly or yearly contract model. The SaaS model sits on top of both the PaaS and the foundational IaaS.

Data as a Service

Because this is a book about big data, we also want you to know about another delivery model called Data as a Service (DaaS). DaaS is closely related to SaaS. DaaS is a platform-independent service that would let you connect to the cloud to store and retrieve your data. In addition, you find a number of specialized data services that are of great benefit in a big data environment. For example, Google offers a service that can process a query with 5 terabytes of data in only 15 seconds. This type of query would typically take ten times as long with a typical data center. Hundreds of specialized analytic services have been developed by companies like IBM and others.

The Cloud as an Imperative for Big Data

Clearly, numerous combinations of deployment and delivery models exist for big data in the cloud. For example, you can utilize a public cloud IaaS or a private cloud IaaS. So, what does this mean for big data and why is the cloud a good fit for it? Well, big data requires distributed clusters of compute power, which is how the cloud is architected. For more on distributed computing, see Chapter 3.

In fact, a number of cloud characteristics make it an important part of the big data ecosystem:

- **Scalability:** Scalability with regard to hardware refers to the capability to go from small to large amounts of processing power with the same architecture. With regard to software, it refers to the consistency of performance per unit of power as hardware resources increase. The cloud can scale to large data volumes. Distributed computing, an integral part of the cloud model, really works on a "divide and conquer" plan. So if you have huge volumes of data, they can be partitioned across cloud servers. An important characteristic of IaaS is that it can dynamically scale. This means that if you wind up needing more resources than expected, you can get them. This ties into the concept of elasticity.
- **Elasticity:** Elasticity refers to the capability to expand or shrink computing resource demand in real time, based on need. One of the benefits of the cloud is that customers have the potential to access as much of a service as they need when they need it. This can be helpful for big data projects where you might need to expand the amount of computing resources you need to deal with the volume and velocity of the data. Of course, this very feature of the cloud that makes it attractive to end users means that the service provider needs to design a platform architecture that is optimized for this kind of service.
- **Resource pooling:** Cloud architectures enable the efficient creation of groups of shared resources that make the cloud economically viable.
- **Self-service:** With self-service, the user of a cloud resource is able to use a browser or a portal interface to acquire the resources needed, say, to run a huge predictive model. This is dramatically different than how you might gain resources from a data center, where you would have to request the resources from IT operations.
- **Often low up-front costs:** If you use a cloud provider, up-front costs can often be reduced because you are not buying huge amounts of hardware or leasing out new space for dealing with your big data. By taking advantage of the economies of scale associated with cloud environments, the cloud can look attractive. Of course, you will need to do your own calculation to evaluate whether you are interested in a public cloud, private cloud, hybrid cloud, or no cloud. We cover this in the section "Where to be careful when using cloud services," later in this chapter.
- **Pay as you go:** A typical billing option for a cloud provider is Pay as You Go (PAYG), which means that you are billed for resources used based on instance pricing. This can be useful if you're not sure what resources you need for your big data project (as long as you don't underbudget).
- **Fault tolerance:** Cloud service providers should have fault tolerance built into their architecture, providing uninterrupted services despite the failure of one or more of the system's components.

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

Making Use of the Cloud for Big Data

Clearly, the very nature of the cloud makes it an ideal computing environment for big data. So how might you use big data together with the cloud? Here are some examples:

- **IaaS in a public cloud:** In this scenario, you would be using a public cloud provider's infrastructure for your big data services because you don't want to use your own physical infrastructure. IaaS can provide the creation of virtual machines with almost limitless storage and compute power. You can pick the operating system that you want, and you have the flexibility to dynamically scale the environment to meet your needs. An example might be using the Amazon Elastic Compute Cloud (Amazon EC2) service, detailed later in the chapter, to run a real-time predictive model that requires data to be processed using massively parallel processing. It might be a service that processes big-box retail data. You might want to process billions of pieces of click-stream data for targeting customers with the right ad in real time.
- **PaaS in a private cloud:** PaaS is an entire infrastructure packaged so that it can be used to design, implement, and deploy applications and services in a public or private cloud environment. PaaS enables an organization to leverage key middleware services without having to deal with the complexities of managing individual hardware and software elements. PaaS vendors are beginning to incorporate big data technologies such as Hadoop and MapReduce into their PaaS offerings. For example, you might want to build a specialized application to analyze vast amounts of medical data. The application would make use of real-time as well as non-real-time data. It's going to require Hadoop and MapReduce for storage and processing. What's great about PaaS in this scenario is how quickly the application can be deployed. You won't have to wait for internal IT teams to get up to speed on the new technologies and you can experiment more liberally. Once you have identified a solid solution, you can bring it in house when IT is ready to support it.
- **SaaS in a hybrid cloud:** Here you might want to analyze "voice of the customer" data from multiple channels. Many companies have come to realize that one of the most important data sources is what the customer thinks and says about their company, their products, and their services. Getting access to voice of the customer data can provide

> invaluable insights into behaviors and actions. Increasingly, customers are “vocalizing” on public sites across the Internet. The value of the customers’ input can be greatly enhanced by incorporating this public data into your analysis. Your SaaS vendor provides the platform for the analysis as well as the social media data. In addition, you might utilize your enterprise CRM data in your private cloud environment for inclusion in the analysis.

TIP

Some industry insiders are using the term *big data applications* when describing applications that run in the cloud that use big data. Examples of this include Amazon.com and LinkedIn. Now some people might argue (and have) that these are really SaaS applications that solve a particular business problem. It’s often a matter of semantics in an emerging space.

Providers in the Big Data Cloud Market

Cloud players come in all shapes and sizes and offer many different products. Some are household names while others are recently emerging. Some of the cloud providers that offer IaaS services that can be used for big data include Amazon.com, AT&T, GoGrid, Joyent, Rackspace, IBM, and Verizon/Terremark.

However, cloud companies and cloud service providers are also offering software targeted specifically for big data. These are described in the following sections.

Amazon’s Public Elastic Compute Cloud

Currently, one of the most high-profile IaaS service providers is Amazon Web Services with its Elastic Compute Cloud (Amazon EC2). Amazon didn’t start out with a vision to build a big infrastructure services business. Instead, the company built a massive infrastructure to support its own retail business and discovered that its resources were underused. Instead of allowing this asset to sit idle, it decided to leverage this resource while adding to the bottom line. Amazon’s EC2 service was launched in 2006 and continues to evolve.

Amazon EC2 offers scalability under the user’s control, with the user paying for resources by the hour. The use of the term *elastic* in the naming of Amazon’s EC2 is significant. Here, elasticity refers to the capability that the EC2 users have to increase or decrease the infrastructure resources assigned to meet their needs.

Amazon also offers other big data services to customers of its Amazon Web Services portfolio. These include the following:

- **Amazon Elastic MapReduce:** Targeted for processing huge volumes of data. Elastic MapReduce utilizes a hosted Hadoop framework (see Chapter 9 for more on Hadoop) running on EC2 and Amazon Simple Storage Service (Amazon S3). Users can now run HBase (a distributed, column-oriented data store).
- **Amazon DynamoDB:** A fully managed not only SQL (NoSQL) database service. DynamoDB is a fault tolerant, highly available data storage service offering self-provisioning, transparent scalability, and simple administration. It is implemented on SSDs (solid state disks) for greater reliability and high performance. We talk more about NoSQL in Chapter 7.
- **Amazon Simple Storage Service (S3):** A web-scale service designed to store any amount of data. The strength of its design center is performance and scalability, so it is not as feature laden as other data stores. Data is stored in "buckets" and you can select one or more global regions for physical storage to address latency or regulatory needs.
- **Amazon High Performance Computing:** Tuned for specialized tasks, this service provides low-latency tuned high performance computing clusters. Most often used by scientists and academics, HPC is entering the mainstream because of the offering of Amazon and other HPC providers. Amazon HPC clusters are purpose built for specific workloads and can be reconfigured easily for new tasks.
- **Amazon RedShift:** Available in limited preview, RedShift is a petabyte-scale data warehousing service built on a scalable MPP architecture. Managed by Amazon, it offers a secure, reliable alternative to in-house data warehouses and is compatible with several popular business intelligence tools.

Google big data services

Google, the Internet search giant, also offers a number of cloud services targeted for big data. These include the following:

- **Google Compute Engine:** A cloud-based capability for virtual machine computing, Google Compute Engine offers a secure, flexible computing environment from energy efficient data centers. Google also offers workload management solutions from several technology partners who have optimized their products for Google Compute Engine.

- **Google Big Query:** Allows you to run SQL-like queries at a high speed against large data sets of potentially billions of rows. Although it is good for querying data, data cannot be modified after it is in it. Consider Google Big Query a sort of Online Analytical Processing (OLAP) system for big data. It is good for ad hoc reporting or exploratory analysis.
- **Google Prediction API:** A cloud-based, machine learning tool for vast amounts of data, Prediction is capable of identifying patterns in data and then remembering them. It can learn more about a pattern each time it is used. The patterns can be analyzed for a variety of purposes, including fraud detection, churn analysis, and customer sentiment. Prediction is covered in more depth in Chapter 12.

Microsoft Azure

Based on Windows and SQL abstractions, Microsoft has productized a set of development tools, virtual machine support, management and media services, and mobile device services in a PaaS offering. For customers with deep expertise in .Net, SQLServer, and Windows, the adoption of the Azure-based PaaS is straightforward.

To address the emerging requirements to integrate big data into Windows Azure solutions, Microsoft has also added Windows Azure HDInsight. Built on Hortonworks Data Platform (HDP), which according to Microsoft, offers 100 percent compatibility with Apache Hadoop, HDInsight supports connection with Microsoft Excel and other business intelligence (BI) tools. In addition to Azure HDInsight can also be deployed on Windows Server.

OpenStack

Initiated by Rackspace and NASA, OpenStack (`www.openstack.org`) is implementing an open-cloud platform aimed at either public or private clouds. While the organization is tightly managed by Rackspace, it moved to a separate OpenStack foundation. Although companies can leverage OpenStack to create proprietary implementations, the OpenStack designation requires conformance to a standard implementation of services.

OpenStack's goal is to provide a massively scaled, multitenant cloud specification that can run on any hardware. OpenStack is building a large ecosystem of partners interested in adopting its cloud platform, including Dell, HP, Intel, Cisco, Red Hat, and IBM, along with at least 100 others that are using OpenStack as the foundation for their cloud offerings. In essence, OpenStack is an open source IaaS initiative built on Ubuntu, an operating system based on the Debian Linux distribution. It can also run on Red Hat's version of Linux.

OpenStack offers a range of services, including compute, object storage, catalog and repository, dashboarding, identity, and networking. In terms of big data, Rackspace and Hortonworks (a provider of an open source data management platform based on Apache Hadoop) announced that Rackspace will release an OpenStack public cloud-based Hadoop service, which will be validated and supported by Hortonworks and will enable customers to quickly create a big data environment.

Where to be careful when using cloud services

Cloud-based services can provide an economical solution to your big data needs, but the cloud has its issues. It's important to do your homework before moving your big data there. Here are some issues to consider:

- **Data integrity:** You need to make sure that your provider has the right controls in place to ensure that the integrity of your data is maintained.
- **Compliance:** Make sure that your provider can comply with any compliance issues particular to your company or industry.
- **Costs:** Little costs can add up. Be careful to read the fine print of any contract, and make sure that you know what you want to do in the cloud.
- **Data transport:** Be sure to figure out how you get your data into the cloud in the first place. For example, some providers will let you mail it to them on media. Others insist on uploading it over the network. This can get expensive, so be careful.
- **Performance:** Because you're interested in getting performance from your service provider, make sure that explicit definitions of service-level agreements exist for availability, support, and performance. For example, your provider may tell you that you will be able to access your data 99.999 percent of the time; however, read the contract. Does this uptime include scheduled maintenance?
- **Data access:** What controls are in place to make sure that you and only you can access your data? In other words, what forms of secure access control are in place? This might include identity management, where the primary goal is protecting personal identity information so that access to computer resources, applications, data, and services is controlled properly.
- **Location:** Where will your data be located? In some companies and countries, regulatory issues prevent data from being stored or processed on machines in a different country.