The two biggest trends in the data center today are cloud computing and big data. This column will examine the intersection of the two. Industry hype has resulted in nebulous definitions for each, so I’ll start by defining terms.

Defining the Cloud

When defining the cloud and big data, it’s helpful to consider both the consumer and producer perspectives. For consumers, the cloud is about consuming hardware or software as a service (SaaS) and the various implications of this approach. For example, pricing models and data governance may change dramatically. In public clouds, the services are run by a third party, while in private clouds, they are owner-operated on premise. Consumers effectively choose the level of vertical integration for their IT; they can choose to own or outsource everything from the data center to the storage, computing, networking, and software infrastructure up to the application.

For producers, on the other hand, the cloud is about the technology that goes into providing service offerings at each level. The technology required to provide an application as a service in the public cloud may differ significantly from the software product that a customer installs to run an internal service. For example, virtual machines are the resource allocation units in most cloud infrastructure offerings, but they might not be used when implementing an application as a public service.

Defining Big Data

For consumers, big data is about using large datasets from new or diverse sources to provide meaningful and actionable information about how the world works. For example, Netflix can use customer data to produce shows tailored to their audiences.

For producers, however, big data is about the technology necessary to handle these large, diverse datasets. Producers characterize big data in terms of volume, variety, and velocity. How much data is there, of what types, and how quickly can you derive value from it?

Although these are good technical descriptions of big data, they don’t fully explain it. Just as adopting a service-oriented approach is the macro trend behind the cloud, there are several macro trends behind big data. The first trend is consumption; we consume data as part of the everyday activities in our personal and working lives. From booking a flight, to finding a partner, to diagnosing disease, data is driving many more decisions today than it has in the past. We live in a relatively new social context where people increasingly want to make data-driven decisions.

Related to consumption, the second trend is instrumentation. We collect data at each step in many of our activities, and much of it is now produced by machines instead of people. From supply chains to Fitbits, we collect information about all our activities with the intent to measure and analyze them.

The third trend is exploration. The relatively easy access to this abundance of data means we can use it to construct, test, and consume experiments that were previously not feasible. Finally, related to
exploration is the concept that the data itself has value. Data is increasingly an asset, not just input to or a byproduct of a business process. This isn’t a new idea of course, but in the context of consumption, instrumentation, and exploration, it’s driving new business models and applications.

Ultimately, big data is about the change in relationship between us and our data and, in the context of this column, the implications of this change on cloud technology.

**Converging Technologies**

So what is the relationship between big data and the cloud? Big data has its origins in the cloud. Apache Hadoop, one of the most widely used big data technologies today, was built on research from Google and initially deployed at Yahoo. Google invented this technology because indexing the Web was infeasible with existing systems. Now companies adopting Hadoop are bringing a cloud architecture into their data centers.

The simultaneous rise of cloud and big data technologies isn’t coincidental—they’re mutually reinforcing. Big data enables the cloud services we consume. For example, SaaS lets us collect data that was infeasible or impossible in a world of packaged software. An application can record every interaction from millions of users. This service in turn drives demand for big data technologies to store, process, and analyze these interactions and inject the value of the analysis back into the application through query and visualization.

The expansion of the cloud continues to drive both the creation of new big data technologies and big data adoption by making it easier and cheaper to access storage and computing resources. Companies can run their big data platforms on infrastructure provided as a service (IaaS) or consume the big data platform as a service (PaaS). Both models work in the public cloud and in on-premise systems.

The decision for enterprises is thus a familiar one: How vertically or horizontally integrated should your infrastructure be? A spectrum of valid options exists, but cloud technology is already enabling more infrastructure outsourcing, whether it’s outsourced to a cloud provider or an internal centralized IT department.

Big data infrastructures also play a role in this trend. For example, recent advances in the Apache Hadoop ecosystem enable more types of workloads and more tenants to share a cluster. What were once discrete systems running on their own hardware are now effectively applications running on Hadoop, sharing the same data and hardware resources. As this abstraction layer evolves and more projects build on it, users will be able to run more types of infrastructures on the same Hadoop cluster, which itself may be running on a cloud infrastructure. As big data infrastructures become more generic, the cloud infrastructure will add more specialized services for data storage, processing, and analysis. Future columns will examine new developments in both areas and the increasing overlap between them.

Another area of exploration for this column will be technologies and trends that are leveraging both cloud computing and big data. The combination of big data, cloud computing, and new algorithms and techniques for visualizing information enables **converged analytics**—performing analytics on data from many different sources. These new techniques for data delivery and data management also enable cloud-based analytics as a service (AaaS). Upcoming columns will cover the development and use of converged analytics and AaaS.

From security and privacy to pricing models, the combination of big data and cloud computing is having a substantial impact on the nontechnical aspects of our lives as well. There is a tension between our desire for converged analytics and cloud computing—which is about sharing more computing resources and data with increasingly diverse tenants—and our desire for better privacy controls and data protection. Usage-based pricing models are forcing us to rethink how we produce and consume technology. Future columns will look at how policies and economics are being shaped by these technological advances.

These are exciting times for both big data and the cloud. Future columns will examine how people are using these trends together, big data developments from cloud builders, and how people are making the cloud better through data. I look forward to exploring all these topics here.***

**ELI COLLINS** is Cloudera’s chief technologist. His research interests include cloud computing and data management. Collins has an MS in computer science from the University of Wisconsin–Madison. Contact him at eli@cloudera.com.