GETTY IMAGES

TIP

An introduction to big data in the cloud

As enterprises add big data projects in the cloud, IT admins need to adjust their skills accordingly. Dive into this comprehensive guide to see what makes a cloud shift so attractive.

Stephen J. Bigelow, Senior Technology Editor

Published: **19 Jul 2021**

Big data is no longer an empty industry buzzword.

Organizations of all sizes recognize the value of data and use it to measure performance, identify challenges and find new opportunities for growth. Big data has also become key in machine learning to train complex models and facilitate AI.

While there is benefit to big data, the sheer amount of computing resources and software services needed to support big data efforts can strain the financial and intellectual capital of even the largest businesses. The cloud has made great strides in filling the need for big data. It can provide almost limitless computing resources and services that make big data initiatives possible for any business.

Here, we will weigh the tradeoffs, evaluate the cloud models and see what services are currently available for big data in the cloud.

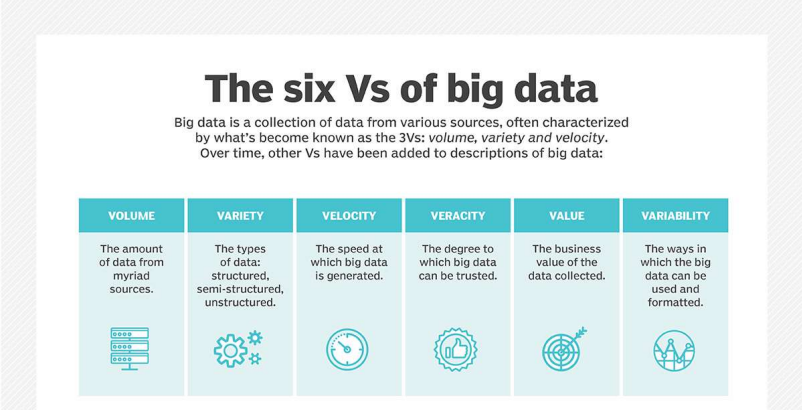
What is big data in the cloud?

Big data and cloud computing are two distinctly different ideas, but the two concepts have become so interwoven that they are almost inseparable. It's important to define the two ideas and see how they relate.

Big data

Big data refers to vast amounts of data that can be structured, semistructured or unstructured. It is all about analytics and is usually derived from different sources, such as user input, IoT sensors and sales data.

Big data also refers to the act of processing enormous volumes of data to address some query, as well as identify a trend or pattern. Data is analyzed through a set of mathematical algorithms, which vary depending on what the data means, how many sources are involved and the business's intent behind the analysis. Distributed computing software platforms, such as Apache Hadoop, Databricks and Cloudera, are used to split up and organize such complex analytics.



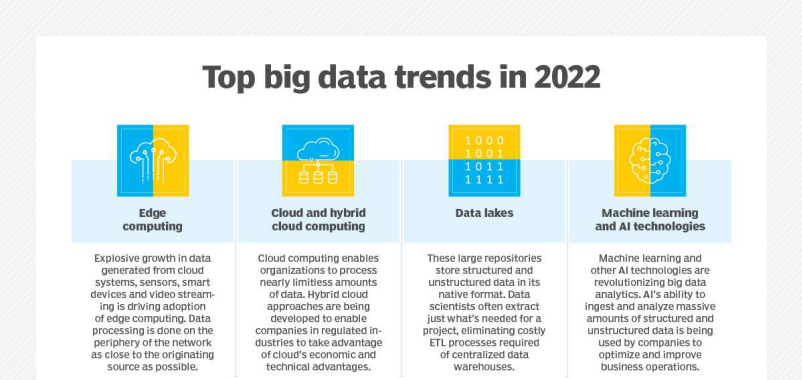
You know the 3 Vs of big data. Here are three more.

The problem with big data is the size of the computing and networking infrastructure needed to build a big data facility. The financial investment in servers, storage and dedicated networks can be substantial, as well as the software knowledge required to set up an effective distributed computing environment. And, once an organization makes an investment in big data, it's only valuable to the business when it's operating -- it's worthless when idle. The demands of big data have long kept the technology limited to only the largest and best-funded organizations. This is where cloud computing has made incredible inroads.

Cloud

Cloud computing provides computing resources and services on demand. A user can easily assemble the desired infrastructure of cloud-based compute instances and storage resources, connect cloud services, upload data sets and perform analyses in the cloud. Users can engage almost limitless resources across the public cloud, use those resources for as long as needed and then dismiss the environment -- paying only for the resources and services that were actually used.

The public cloud has emerged as an ideal platform for big data. A cloud has the resources and services that a business can use on demand, and the business doesn't have to build, own or maintain the infrastructure. Thus, the cloud makes big data technologies accessible and affordable to almost any size of enterprise.



Get to know 2021's top big data trends.

The pros of big data in the cloud

The cloud brings a variety of important benefits to businesses of all sizes. Some of the most immediate and substantial benefits of big data in the cloud include the following.

Scalability

A typical business data center faces limits in physical space, power, cooling and the budget to purchase and deploy the sheer volume of hardware it needs to build a big data infrastructure. By comparison, a public cloud manages hundreds of thousands of servers spread across a fleet of global data centers. The infrastructure and software services are already there, and users can assemble the infrastructure for a big data project of almost any size.

Agility

Not all big data projects are the same. One project may need 100 servers, and another project might demand 2,000 servers. With cloud, users can employ as many resources as needed to accomplish a task and then release those resources when the task is complete.

Cost

A business data center is an enormous capital expense. Beyond hardware, businesses must also pay for facilities, power, ongoing maintenance and more. The cloud works all those costs into a flexible rental model where resources and services are available on demand and follow a pay-per-use model.

Accessibility

Many clouds provide a global footprint, which enables resources and services to deploy in most major global regions. This enables data and processing activity to take place proximally to the region where the big data task is located. For example, if a bulk of data is stored in a certain region of a cloud provider, it's relatively simple to implement the resources and services for a big data project in that specific cloud region -- rather than sustaining the cost of moving that data to another region.

Resilience

Data is the real value of big data projects, and the benefit of cloud resilience is in data storage reliability. Clouds replicate data as a matter of standard practice to maintain high availability in storage resources, and even more durable storage options are available in the cloud.

The cons of big data in the cloud

Public clouds and many third-party big data services have proven their value in big data use cases. Despite the benefits, businesses must also consider some of the potential pitfalls. Some major disadvantages of big data in the cloud can include the following.

Network dependence

Cloud use depends on complete network connectivity from the LAN, across the internet, to the cloud provider's network. Outages along that network path can result in increased latency at best or complete cloud inaccessibility at worst. While an outage might not impact a big data project in the same ways that it would affect a mission-critical workload, the effect of outages should still be considered in any big data use of the cloud.

Storage costs

Data storage in the cloud can present a substantial long-term cost for big data projects. The three principal issues are data storage, data migration and data retention. It takes time to load large amounts of data into the cloud, and then those storage instances incur a monthly fee. If the data is moved again, there may be additional fees. Also, big data sets are often time-sensitive, meaning that some data may have no value to a big data analysis even hours into the future. Retaining unnecessary data costs money, so businesses must employ comprehensive data retention and deletion policies to manage cloud storage costs around big data.

Be wary of these 10 big data challenges.

Security

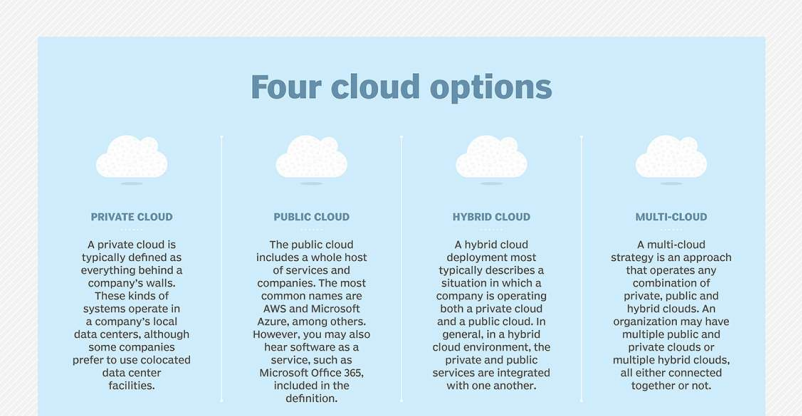
The data involved in big data projects can involve proprietary or personally identifiable data that is subject to data protection and other industry- or government-driven regulations. Cloud users must take the steps needed to maintain security in cloud storage and computing through adequate authentication and authorization, encryption for data at rest and in flight, and copious logging of how they access and use data.

Lack of standardization

There is no single way to architect, implement or operate a big data deployment in the cloud. This can lead to poor performance and expose the business to possible security risks. Business users should document big data architecture along with any policies and procedures related to its use. That documentation can become a foundation for optimizations and improvements for the future.

Choose the right cloud deployment model

So, which cloud model is ideal for a big data deployment? Organizations typically have four different cloud models to choose from: public, private, hybrid and multi-cloud. It's important to understand the nature and tradeoffs of each model.



Which deployment model is right for you?

Private cloud

Private clouds give businesses control over their cloud environment, often to accommodate specific regulatory, security or availability requirements. However, it is more costly because a business must own and operate the entire infrastructure. Thus, a private cloud might only be used for sensitive small-scale big data projects.

Public cloud

The combination of on-demand resources and scalability makes public cloud ideal for almost any size of big data deployment. However, public cloud users must manage the cloud resources and services it uses. In a shared responsibility model, the public cloud provider handles the security of the cloud, while users must configure and manage security in the cloud.

Hybrid cloud

A hybrid cloud is useful when sharing specific resources. For example, a hybrid cloud might enable big data storage in the local private cloud -- effectively keeping data sets local and secure -- and use the public cloud for compute resources and big data analytical services. However, hybrid clouds can be more complex to build and manage, and users must deal with all of the issues and concerns of both public and private clouds.

Multi-cloud

With multiple clouds, users can maintain availability and use cost benefits. However, resources and services are rarely identical between clouds, so multiple clouds are more complex to manage. This cloud model also has more risks of security oversights and compliance breaches than single public cloud use. Considering the scope of big data projects, the added complexity of multi-cloud deployments can add unnecessary challenges to the effort.

Review big data services in the cloud

While the underlying hardware gets the most attention and budget for big data initiatives, it's the services -- the analytical tools -- that make big data analytics possible. The good news is that organizations seeking to implement big data initiatives don't need to start from scratch.

Providers not only offer services and documentation, but can also arrange for support and consulting to help businesses optimize their big data projects. A sampling of available big data services from the top three providers include the following.

AWS

- Amazon Elastic MapReduce
- AWS Deep Learning AMIs
- Amazon SageMaker

Microsoft Azure

- Azure HDInsight
- Azure Analysis Services
- Azure Databricks

Google Cloud

- Google BigQuery
- Google Cloud Dataproc
- Google Cloud AutoML

Keep in mind that there are numerous capable services available from third-party providers. Typically, these providers offer more niche services, whereas major providers follow a one-size-fits-all strategy for their services. Some third-party options include the following:

- Cloudera

- Hortonworks Data Platform
- Oracle Big Data Service
- Snowflake Data Cloud

Next Steps

What a big data strategy is and how to build one

17 top big data tools and technologies to know about

Dig Deeper on Cloud app development and management

big data as a service (BDaaS)

By: Craig Stedman

The case against data center consolidation

By: Stephen Bigelow

Decision points in storage for artificial intelligence, machine learning and big data

By: Stephen Pritchard

4 reasons cloud data repatriation is happening

By: Paul Kirvan

Data Center

Quantum data centers might be the way of the future

Quantum computing has lots of potential for high compute applications. But the technology is still in the early stages, so it may...

Learn different data lake vs. data warehouse uses

Data lakes and data warehouses both store big data. When choosing a lake or warehouse, consider factors such as cost and what ...