BIG DATA IN THE CLOUD

# How to Overcome the Top Five Big Data Challenges – It's All About the Cloud


## Summary

As easy as it is to get swept up by the hype surrounding big data, it's just as easy for organizations to become discouraged by the challenges they encounter while implementing a big data initiative. Concerns regarding big data skill sets (and the lack thereof), security, the unpredictability of data, unsustainable costs, and the need to make a business case can bring a big data initiative to a screeching halt.

However, given big data's power to transform business, it's critical that organizations overcome these chal - lenges and realize the value of big data. The cloud can help organizations to do so. Drawing from IDG's 2015 Big Data and Analytics Survey , this *white paper analyzes the top five* challenges companies face when undergoing a big data initiative and *explains how they can effectively* overcome them in the cloud.


## Introduction

It's impossible to turn left or right these days without hearing the words "big data" or "cloud"—and for good reason. In order to be competitive in today's marketplace, companies must make informed business decisions that will produce real results, whether those results are increasing revenue, retaining customers, or improving product quality. Big data is a key enabler to making those goals happen.

IDG defines big data as "large volumes of a wide variety of data collected from various sources across the enterprise, including transactional data from enterprise applications/databases, social media data, mobile device data, unstructured data/ documents, machine-generated data, and more." IDG goes on to say, "Data in high volume, high velocity, and from a high variety of information assets can deliver enhanced insights and decision making."

Big data enables organizations to understand their business on a deeper level and make strategic decisions in real time. In fact, 1 in 3 respondents to IDG's 2015 Big Data and Analytics Survey report improved quality of decision making, and better planning and forecasting as a result of their big data initiatives .

But as with any new technology, there are challenges. The first challenge: data volume and velocity. Bigger data that's changing in real time means existing tools and approaches won't work. Consider also the sources: Big data comes at you from, in some cases, literally millions of places—customers, sensors, websites, social media. You get the idea.

The old way of approaching this would be to grow your capacity by building or expanding to handle big data workloads. That's a resource-intense move—it's expensive, time-consuming, requires lots of IT staff time and skills, and doesn't allow your business to move fast enough. You can end up spending more time and money on infrastructure than building great products and solutions.

The cloud can help with a lot of these problems. It comes as no surprise that cloud and predictive analytics are among the top three disruptive technologies most likely to have an impact on organizations in the next three to five years. (The third is self-service IT.)
If you're going to leverage predictive analytics from big data, the cloud can be a key enabler with many advantages.

Examples of successful big data initiatives are numerous. Consider this one involving Major League Baseball Advanced Media (MLBAM). The organization wanted to develop a new way to capture and analyze every play using data collection and analysis tools in order to provide a more engaging experience for a younger generation of fans. To do that, MLBAM built the Player Tracking System, which collects real-time game data and delivers valuable insights to fans, thus shifting from historical/static analysis to real-time analytics during the game. Not only has the system increased fan engagement, it has also presented opportunities for new streams of revenue, as other sports leagues are interested in acquiring MLBAM's services.

## Top five big data challenges

Despite the numerous success stories, embarking on a big data initiative isn't necessarily easy. In fact, it presents a number of challenges, any of which can derail a project before it even gets started. In its Big Data and Analytics Survey, IDG identifies these five top challenges:

1. Skills shortage

The big data ecosystem is moving fast—so fast that it's nearly impossible to keep up. New tools, capabilities, and frameworks evolve and mature in a matter of months, resulting in a skills gap that can easily impede a big data initiative. In fact, 48% of organizations cite the shortage of employees with data analysis and data management skills as their No. 1 big data challenge. The demand for big data skills—especially in analytics—is so great that 70% of respondents are planning to hire for big data skill sets in the next 12 to 18 months.

Organizations are questioning how they can make long-term IT investments, leverage existing skills, and obtain new ones. There's also the issue of knowing which technologies and frameworks—Hive, Pig, MapReduce, Spark, NoSQL stores, etc.—are best for any given big data initiative. Meanwhile, the strong demand for high analytical and data management skills continues to grow.

Leveraging the cloud enables organizations to tap the latest technologies, without committing substantial time and resources to ongoing setup, maintenance, and upgrades. The cloud also allows organizations to use the skills they already have, while managed services can supplement skills they need.

2. Cost

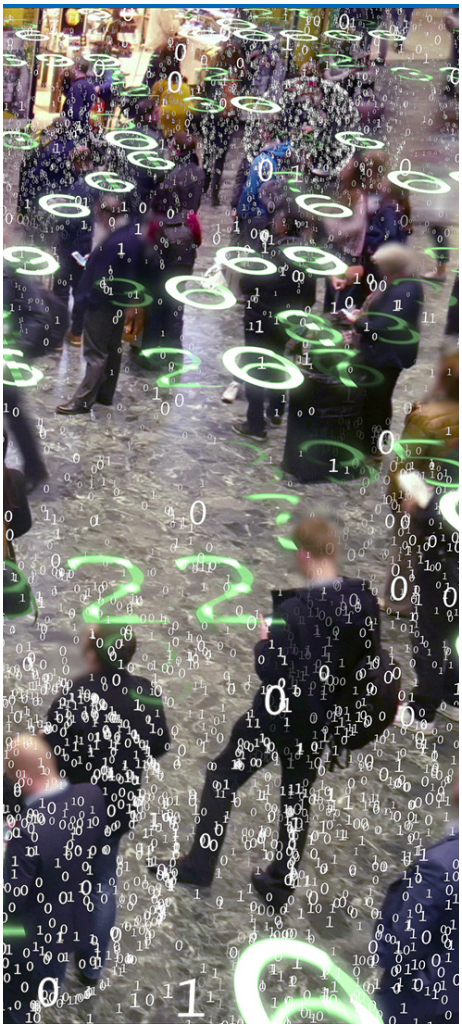
Cited by 47% of respondents, budgetary limitations are the second biggest challenge facing companies today as they embark on a big data initiative. The fact that cost is the No. 1 concern two years in a row prior is a testament to this challenge.

Most big data technologies require large clusters of servers that entail long provisioning and setup cycles, resulting in significant capital expenditures and maintenance overhead. To make matters more complicated, the growing volume, variety, and velocity of data from either existing applications or new business requirements can result in unsustainable IT costs. Organizations need to know how to get value from big data without breaking the bank. They must be able to scale the infrastructure to manage big data while still reducing IT costs. That is exactly what the cloud enables organizations to do. The cloud eliminates the need to procure and maintain hardware and software infrastructure—and the large capital expenditures that go with them—and instead reallocate dollars into core innovation instead of just keeping the lights on.

3. The unpredictability of data

Big data comes from a wide variety of sources, from legacy applications and transactional systems, to machine-generated data, mobile devices, web logs, and social media. This makes it even more difficult and inefficient to predict the required capacity. A single event can cause sudden changes in data volumes and workloads. For example, a financial services organization can experience volume fluctuations by a factor of 10 on any given day, depending on market conditions.

One-quarter of organizations are challenged by big data's growing demands on storage capacity/infrastructure. Not only do organizations have to size their infrastructure, they must also determine how they'll easily scale to address fluctuating storage and computing requirements. It's inefficient and cost-prohibitive for almost any



Even big data security challenges can be addressed in the cloud by choosing a provider that has robust controls in place for data privacy and security.

organization to size its infrastructure to support 10x volumes and let that extra capacity sit unused for 90% of the time. Additional issues include escalating infrastructure and maintenance costs due to data growth as well as ensuring adequate bandwidth to support innovation through experimentation, plus data capture and analysis.

In the cloud, there's no need to size an infrastructure for maximum capacity. Its elastic properties allow organizations to dynamically scale the infrastructure up or down as needed.

4. Security

As organizations collect, store, and analyze increasing amounts of data from new and existing sources, security becomes of greater concern. Nearly 35% of survey respondents either aren't sure or don't think that their existing security solutions and products provide adequate data security. They struggle to control data access, secure data assets, and protect the infrastructure. Ultimately, they're left to determine how to ensure compliance, governance, and security without compromising on agility and performance. The healthcare industry, for example, must meet HIPAA compliance—a complex undertaking that involves securing not just data but access to data via computers, printers, and copiers.

The financial services industry also illustrates the challenges. Big data is one of the most promising new technologies to evolve recently, focused around improving customer intelligence, reducing risk, and meeting regulatory compliance requirements. Take the big data challenges we've discussed, like security, and consider:

- The strict governance and compliance requirements already in place for financial information.
- The fact that essentially all data created or used by a financial services firm is regulated, potentially sensitive, or private.
- The alphabet soup of regulations—GLBA, SOX/J-SOX, MiFID II, Basel II, even the USA Patriot Act—that financial services organizations must address.

Big data also implies that your information isn't at rest: The point is using that data for insights that lead to better business outcomes. That data is continuously generated, processed, and analyzed by multiple users and systems.

Combine that with day-to-day security needs like safeguarding against insider and outsider attacks, and assuring the security and privacy of thousands of customers' data, and the scope of the challenge becomes clearer.

Even big data security challenges can be addressed in the cloud by choosing a provider that has robust controls in place for data privacy and security. In fact, it's not unusual for the cloud to be more secure than the corporate data center. Since cloud service providers are in the business of offering robust compute infrastructure, it's in their best interest to maintain a secure environment. To that end, many cloud providers have accumulated best practices and experience from multiple organizations with some of the most stringent security requirements.

5. Making a business case

In many cases it's up to IT to make the business case for big data. According to IDG, IT heads are significantly more likely than non-IT heads to be in charge of determining requirements and the business need for a solution. They recommend and select vendors, approve and authorize purchase, and sell the solution outside the IT team. But business leaders aren't siting on the sidelines. IDG Research shows 45% of organizations

The fact of the matter is, big data is still relatively immature and can involve a degree of experimentation, the cost of which can be very high.

report their CEO is involved with data-driven initiatives. The CFO and line-of-business executives also increasingly play key roles in big data projects.

If you haven't built a solid business case, and gathered input from powerful allies such as key business stakeholders, then chances are you're not going to get approval for the resources you need.

It's no wonder that the ability to demonstrate ROI is a challenge cited by a quarter of all enterprises. The fact of the matter is, big data is still relatively immature and can involve a degree of experimentation, the cost of which can be very high. In order to run experiments on specific initiatives, organizations must do the undifferentiated heavy lifting that translates into a lot of time and effort. This slows down the pace of innovation and ultimately lowers the value of a big data initiative.

In many cases, the easiest way to demonstrate ROI is to lower the total cost of ownership, all other things being equal. But as mentioned previously, the cost of managing big data with traditional infrastructures is unsustainable. Rearchitecting existing workloads using the cloud can help reduce costs significantly. In addition, leveraging the cloud can help accelerate the pace of innovation by reducing the cost of experimentation. Successful experiments will show measurable benefits that, once in place, will spark the need for more.

## Using the cloud to overcome these challenges

The right cloud approach can help minimize, and in some cases even eliminate, many of the barriers to deploying big data applications. Like big data, the cloud is a highly disruptive force that's transforming the way organizations operate and do business. However, when combined, the impact of cloud and big data is even greater.

But just deciding to leverage the cloud isn't going to solve your big data problems overnight. The problem is, lots of cloud providers offer just a subset of everything you need. And, what they do have requires a lot of integration work—often leaving you with a big engineering problem and forcing difficult trade-offs: Price or scalability? Performance or ease of use? Agility or security?

When evaluating cloud providers, look for one that can directly address each of thsee challenges.

- Skills shortage: You need broad capabilities to build, scale, and securely deploy big data applications. These capabilities should cover all the different aspects of big data, from data collection, to storage, analytics, and data visualization. Look for a cloud provider that offers managed services that minimize the administrative overhead, and that is compatible with a breadth of technologies in the big data space. This will allow you to leverage the skills you have and get help for those you don't.
- Cost and making a business case: Moving to the cloud will eliminate the need to procure and maintain hardware. To help build the business case, select a provider that enables you to lower TCO. Flexible pricing models—from reserved instances, to on-demand capacity, and even spot instances—can provide huge savings opportunities to lower the cost structure of managing and processing your data.
- The unpredictability of data: Your cloud provider should allow you to scale up or down quickly and easily to respond to changes in demand.

aws | intel

AdRoll, an industry leader in ad retargeting, turned to the cloud following a period of rapid business growth. In order to continue to serve up ads *effectively, the company needed the flexibility to add capacity* at a moment's notice. High performance and automation were also necessary to ensure that the system could quickly respond to bids.

By moving its core systems to Amazon Web Services (AWS), AdRoll not only reduced costs but also gained the ability to *handle incoming traffic from* Facebook, Google, Yahoo, and other popular sites, so that it can serve up more than 50 billion impressions a day. AWS allowed AdRoll to scale and optimize its algorithms—and get rid of extra capacity—all the while saving time and money.

For instance, decoupling storage from computing capacity allows organizations to select only the type and size of resources they need and pay only for what they use.

- Compute options: Your provider should offer a deep and varied set of compute options appropriate for the widest range of big data workloads. These include compute-optimized instances; GPU instances for high-performance computation; memory-optimized instances with terabytes of RAM for memory-intensive applications; and storage-optimized instances with very fast SSD storage for massively parallel data warehousing applications, Hadoop, or NoSQL databases.

- Security: Look for a cloud infrastructure that is designed to be secure and constantly audited for compliance with a variety of industry standards such as HIPAA, PCI DSS, or FedRAMP. Make sure the cloud provider offers audit-friendly services and compliance programs to help you meet your security and governance requirements. And be sure the provider offers data encryption at rest and in transit for all services as well as a broad set of encryption options for data.

The cloud's very nature makes it well suited for big data. Due to its scalability, elasticity, and economic model, the cloud allows organizations to scale up and down as needed without building out—and investing in—an environment sized for peak capacity. The cloud enables organizations to reduce the costs associated with the heavy lifting and instead reinvest the savings in projects that deliver value to the organization. Measurable savings will help gain more sponsors, and those savings can then be used to fund other big data initiatives.

## Amazon Web Services: Your big data partner

Big data entails much more than collecting large volumes of structured, semi structured, and unstructured data—that's just the beginning. In order to derive valuable insights from big data it must also be securely stored, cleansed, aggregated, sorted, joined, analyzed, etc. That's why organizations need Amazon Web Services. AWS offers a complete set of on-demand cloud services for big data. AWS provides a broad and deep set of managed services for big data analytics in the market. They are also low cost, scalable, high performance, easy to use, and secure. AWS' comprehensive set of capabilities covers the entire range of approaches to big data analytics, including big data stores, data warehousing, distributed analytics (supporting Hadoop, Spark, HBase, Hive, Pig, and Yarn), machine learning, and business intelligence.

With AWS, there's no hardware to procure and no infrastructure to maintain and scale. A self-service model provides on-demand access to compute, storage, and networking capacity. You have the flexibility to use managed services where provisioning, availability, durability, recovery, and backup services are done for you, or you can build and deploy your own platform using popular tools, including open source tools for big data. Plus, you can ingest data at any velocity, from any variety of sources, and process and analyze data in near real time with tools you're already familiar with.

With the infrastructure out of the way, organizations can focus on their core business competencies. They can easily and efficiently test new ideas so that you can quickly realize the value of your big data initiatives. AWS provides a flexible and open environment that gives you the

*AWS provides a flexible and* open environment that gives you the ability to develop your own tools, install your own software, procure managed services, or use partner solutions.

ability to develop your own tools, install your own software, procure managed services, or use partner solutions.

AWS is a secure environment trusted by companies across a variety of industries—including those in highly regulated segments like financial, healthcare, and government—to run their big data applications. AWS environments are continuously audited with certifications from accreditation bodies across geographies and verticals. These cover HIPAA, PCI DSS, ISO 9001:2008, ISO 27018:2014, FedRAMP, and Cloud Security Alliance, to name just a few. Governance-focused, audit-friendly service features are combined with applicable compliance and audit standards to help organizations establish and operate in an AWS security control environment.

By migrating to AWS, FINRA—the Financial Industry Regulatory Authority—has created a flexible platform that can adapt to changing market dynamics while providing its analysts with the tools to interactively query multi-petabyte data sets. FINRA is dedicated to investor protection and market integrity. It regulates one critical part of the securities industry—brokerage firms doing business with the public in the United States. To respond to rapidly changing market dynamics, FINRA moved about 90 percent of its data volumes to Amazon Web Services, using AWS to capture, analyze, and store a daily influx of 37 billion records.

Building and running your big data applications in the cloud can provide the scale and performance to quickly discover business insights. AWS provides high-performance, on-demand Intel® Xeon® processors so that you can be sure you're getting the latest hardware without having to invest in new servers, build a cluster, or schedule time on your existing infrastructure. In addition, AWS and Intel® offer a variety of compute options to allow you to find the right balance of price and performance for your application.

Powered by Intel® Xeon® processors, AWS provides a broad set of elastic compute instances that are ideal for big data workloads. Organizations can easily provision resizable compute capacity in the cloud, including dense storage for data warehousing and Hadoop computing; memory optimized for in-memory analytics; compute optimized for science and engineering apps; and instances for GPU workloads. In addition, AWS lets you bid on spare computing capacity. By choosing the price you're willing to pay per instance, per hour, you can reduce costs while growing your application's compute capacity.

## Conclusion

Going forward, big data will play an increasingly important role in enabling organizations to make smarter, faster business decisions. But organizations don't have to be held back by a skills shortage, cost, the unpredictability of data, security issues, or the difficulty of creating a business case. The cloud, particularly AWS, can address many of these requirements. AWS enables organizations to iterate on big data analytics and focus on business needs without worrying about the IT infrastructure needed to collect, store, and process big data. With AWS solutions, powered by Intel, organizations can analyze data faster and at a lower cost, to get to their next "ah-ha" moment sooner.

To learn more about Amazon's **big data offerings, visit the following sites:**

https://aws.amazon.com/big-data

- https://aws.amazon.com/intel