Security Compliance Auditing and Protection

PROTECTING BIG DATA ANALYTICS

It is sad to report that protecting data is an often forgotten inclination in the data center, an afterthought that falls behind current needs. The launch of Big Data initiatives is no exception in the data center, and protection is too often an afterthought. Big Data offers more of a challenge than most other data center technologies, making it the perfect storm for a data protection disaster.

The real cause of concern is the fact that Big Data contains all of the things you don't want to see when you are trying to protect data. Big Data can contain very unique sample sets—for example, data from devices that monitor physical elements (e.g., traffic, movement, soil pH, rain, wind) on a frequent schedule, surveillance cameras, or any other type of data that are accumulated frequently and in real time. All of the data are unique to the moment, and if they are lost, they are impossible to recreate.

That uniqueness also means you cannot leverage time-saving backup preparation and security technologies, such as deduplication; this greatly increases the capacity requirements for backup subsystems, slows down security scanning, makes it harder to detect data corruption, and complicates archiving. There is also the issue of the large size and number of files often found in

Big Data analytic environments. In order for a backup application and associated appliances or hardware to churn through a large number of files, bandwidth to the backup systems and/or the backup appliance must be large, and the receiving devices must be able to ingest data at the rate that the data can be delivered, which means that significant CPU processing power is necessary to churn through billions of files.

There is more to backup than just processing files. Big Data normally includes a database component, which cannot be overlooked. Analytic information is often processed into an Oracle, NoSQL, or Hadoop environment of some type, so real-time (or live) protection of that environment may be required. A database component shifts the backup ideology from a massive number of small files to be backed up to a small number of massive files to be backed up. That changes the dynamics of how backups need to be processed.

BIG DATA AND COMPLIANCE

Compliance issues are becoming a big concern in the data center, and these issues have a major effect on how Big Data is protected, stored, accessed, and archived. Whether Big Data is going to reside in the data warehouse or in some other more scalable data store remains unresolved for most of the industry; it is an evolving paradigm. However, one thing is certain: Big Data is not easily handled by the relational databases that the typical database administrator is used to working with in the traditional enterprise database server environment. This means it is harder to understand how compliance affects the data.

Big Data is transforming the storage and access paradigms to an emerging new world of horizontally scaling, unstructured databases, which are better at solving some old business problems through analytics. More important, this new world of file types and data is prompting analysis professionals to think of new problems to solve, some of which have never been attempted before. With that in mind, it becomes easy to see that a rebalancing of the database landscape is about to commence, and data architects will finally embrace the fact that relational databases are no longer the only tool in the tool kit. This has everything to do with compliance. New data types and methodologies are still expected to meet the legislative requirements placed on businesses by compliance laws. There will be no excuses accepted and no passes given if a new data methodology breaks the law.

The lessons learned to show that there is away to keep Big Data secure and in compliance. A combination of technologies has been assembled to meet four important goals:

1. **Control access by process, not job function**:

Server and network administrators, cloud administrators, and other employees often have access to more information than their jobs require because the systems simply lack the appropriate access controls. Just because a user has operating system–level access to a specific server does not mean that he or she needs, or should have, access to the Big Data stored on that server.

2. **Secure the data at rest:** Most consumers today would not conduct an online transaction without seeing the familiar padlock symbol or at least a certification notice designating that particular transaction as encrypted and secure. So why wouldn't you require the same data to be protected at rest in a Big Data store? All Big Data, especially sensitive information, should remain encrypted, whether it is stored on a disk, on a server, or in the cloud and regardless of whether the cloud is inside or outside the walls of your organization.

3. **Protect the cryptographic keys and store them separately from the data**.

Cryptographic keys are the gateway to the encrypted data. If the keys are left unprotected, the data are easily compromised. Organizations—often those that have cobbled together their own encryption and key management solution—will sometimes leave the key exposed within the configuration file or on the very server that stores the encrypted data. This leads to the frightening reality that any user with access to the server, authorized or not, can access the key and the data. In addition, that key may be used for any number of other servers. Storing the cryptographic keys on a separate, hardened server, either on the premises or in the cloud, is the best practice for keeping data safe and an important step in regulatory compliance. The bottom line is to treat key security with as much, if not greater, rigor than the data set itself.

4. **Create trusted applications and stacks to protect data from rogue users**.

You may encrypt your data to control access, but what about the user who has access to the configuration files that define the access controls to those data? Encrypting more than just the data and hardening the security of your overall environment—including applications, services, and configurations—gives you peace of mind that your sensitive information is protected from malicious users and rogue employees. There is still time to create and deploy appropriate security rules and compliance objectives. The health care industry has helped to lay some of the groundwork. However, the slow development of laws and regulations works in favor of those trying to get ahead on Big Data. Currently, many of the laws and regulations have not addressed the unique challenges of data warehousing. Many of the regulations do not address the rules for protecting data from different customers at different levels. Similarly, social media applications that are collecting tons of unregulated yet potentially sensitive data may not yet be a compliance concern. But they are still a security problem that if not properly addressed now may be regulated in the future. Social networks are accumulating massive amounts of unstructured data—a primary fuel for Big Data, but they are not yet regulated, so this is not a compliance concern but remains as a security concern.

There are still some very basic rules that should be used to enable security while not derailing the value of Big Data:

- **Ensure that security does not impede performance or availability**.

Big Data is all about handling volume while providing results, being able to deal with the velocity and variety of data, and allowing organizations to capture, analyze, store, or move data in real time. Security controls that limit any of these processes are a nonstarter for organizations serious about Big Data.

- **Pick the right encryption scheme**.

Some data security solutions encrypt at the file level or lower, such as including specific data values, documents, or rows and columns. Those methodologies can be cumbersome, especially for key management. File level or internal file encryption can also render data unusable because many applications cannot analyze encrypted data. Likewise, encryption at the operating system

level, but without advanced key management and process based access controls, can leave Big Data woefully insecure. To maintain the high levels of performance required to analyze Big Data, consider a transparent data encryption solution optimized for Big Data.

- **Ensure that the security solution can evolve with your changing requirements.**

Vendor lock-in is becoming a major concern for many enterprises. Organizations do not want to be held captive to a sole source for security, whether it is a single server vendor, a network vendor, a cloud provider, or a platform. The flexibility to migrate between cloud providers and models based on changing business needs is a requirement, and this is no different with Big Data technologies. When evaluating security, you should consider a solution that is platform-agnostic and can work with any Big Data file system or database, including Hadoop, Cassandra, and MongoDB..

Best Practices for Big Data Analytics

The evolutionary aspect of Big Data tends to affect best practices, so what may be best today may not necessarily be best tomorrow. That said, there are still some core proven techniques that can be applied to Big Data analytics and that should withstand the test of time. With new terms, new skill sets, new products, and new providers, the world of Big Data analytics can seem unfamiliar, but tried and- true data management best practices do hold up well in this still emerging discipline. As with any business intelligence (BI) and/or data warehouse initiative, it is critical to have a clear understanding of an organization's data management requirements and a well-defined strategy before venturing too far down the Big Data analytics path. Big Data analytics is widely hyped, and companies in all sectors are being flooded with new data sources and ever larger amounts of information. Yet making a big investment to attack the Big Data problem without first figuring out how doing so can really add value to the business is one of the most serious missteps for would-be users.

START SMALL WITH BIG DATA

When analyzing Big Data, it makes sense to define small, high-value opportunities and use those as a starting point. Ideally, those smaller tasks will build the expertise needed to deal with the larger questions an organization may have for the analytics process. As companies expand the data sources and types of information they are looking to analyze, and as they start to create the all-important analytical models that can help them uncover patterns and correlations in both structured and unstructured data, they need to be vigilant about homing in on the findings that are most important to their stated business objectives. It is critical to avoid situations in which you end up with a process that identifies news patterns and data relationships that offer little value to the business process. That creates a dead spot in an analytics matrix where patterns, though new, may not be relevant to the questions being asked.

Successful Big Data projects tend to start with very targeted goals and focus on smaller data sets. Only then can that success be built upon to create a true Big Data analytics methodology that starts small and grows after the practice has served the enterprise rather well, allowing value to be created with little upfront investment while preparing the company for the potential windfall of information that can be derived from analytics. That can be accomplished by starting with "small bites" (i.e., taking individual data flows and migrating those into different systems for converged processing). Over time, those small bites will turn into big bites, and Big Data will be born. The ability to scale will prove important—as data collection increases, the scale of the system will need to grow to accommodate the data.

THINKING BIG

Leveraging open source Hadoop technologies and emerging packaged analytics tools makes an open source environment more familiar to business analysts trained in using SQL. Ultimately, scale will become the primary factor when mapping out a Big Data analytics road map, and business analysts will need to eschew the ways of SQL to grasp the concept of distributed platforms that run on nodes and clusters. It is critical to consider what the buildup will look like. It can be accomplished by determining how much data will need to be gathered six months from

now and calculating how many more servers may be needed to handle it. You will also have to make sure that the software is up to the task of scaling. One big mistake is to be ignorant about the potential growth of the solution and the potential popularity of the solution once it is rolled into production.

AVOIDING WORST PRACTICES

There are many potential reasons that Big Data analytics projects fall short of their goals and expectations, and in some cases it is better to know what not to do rather than knowing what to do. This leads us to the idea of identifying "worst practices," so that you can avoid making the same mistakes that others have made in the past. It is better to learn from the errors of others than to make your own. Some worst practices to look out for are the following:

- **Thinking "If we build it, they will come."**

Many organizations make the mistake of assuming that simply deploying a data warehousing or BI system will solve critical business problems and deliver value. However, IT as well as BI and analytics program managers get sold on the technology hype and forget that business value is their first priority; data analysis technology is just a tool used to generate that value. Instead of blindly adopting and deploying something, Big Data analytics proponents first need to determine the business purposes that would be served by the technology in order to establish a business case—and only then choose and implement the right analytics tools for the job at hand. Without a solid understanding of business requirements, the danger is that project teams will end up creating a Big Data disk farm that really isn't worth anything to the organization, earning the teams an unwanted spot in the "data doghouse."

- **Assuming that the software will have all of the answers**.

Building an analytics system, especially one involving Big Data, is complex and resource-intensive. As a result, many organizations hope the software they deploy will be a magic bullet

that instantly does it all for them. People should know better, of course, but they still have hope. Software does help, sometimes dramatically. But Big Data analytics is only as good as the data being analyzed and the analytical skills of those using the tools.

- **Not understanding that you need to think differently.**

Insanity is often defined as repeating a task and expecting different results, and there is some modicum of insanity in the world of analytics. People forget that trying what has worked for them in the past, even when they are confronted with a different situation, leads to failure. In the case of Big Data, some organizations assume that big just means more transactions an large data volumes. It may, but many Big Data analytics initiatives involve unstructured and semi structured information that needs to be managed and analyzed in fundamentally different ways than is the case with the structured data in enterprise applications and data warehouses. As a result, new methods and tools might be required to capture, cleanse, store, integrate, and access at least some of your Big Data.

- **Forgetting all of the lessons of the past.**

Sometimes enterprises go to the other extreme and think that everything is different with Big Data and that they have to start from scratch. This mistake can be even more fatal to a Big Data analytics project's success than thinking that nothing is different. Just because the data you are looking to analyze are structured differently doesn't mean the fundamental laws of data management have been rewritten.

- **Not having the requisite business and analytical expertise**.

A corollary to the misconception that the technology can do it all is the belief that all you need are IT staffers to implement Big Data analytics software. First, in keeping with the theme mentioned earlier of generating business value, an effective Big Data analytics program has to incorporate extensive business and industry knowledge into both the system design stage and ongoing operations. Second, many organizations underestimate the extent of the analytical skills that are needed. If Big Data analysis is only about building reports and dashboards, enterprises can probably just leverage their existing BI expertise. However, Big Data analytics typically

involves more advanced processes, such as data mining and predictive analytics. That requires analytics professionals with statistical, actuarial, and other sophisticated skills, which might mean new hiring for organizations that are making their first forays into advanced analytics.

- **Treating the project like a science experiment**.

Too often, companies measure the success of Big Data analytics programs merely by the fact that data are being collected and then analyzed. In reality, collecting and analyzing the data is just the beginning. Analytics only produces business value if it is incorporated into business processes, enabling business managers and users to act on the findings to improve organizational performance and results. To be truly effective, an analytics program also needs to include a feedback loop for communicating the success of actions taken as a result of analytical findings, followed by a refinement of the analytical models based on the business results.

- **Promising and trying to do too much.**

Many Big Data analytics projects fall into a big trap: Proponents oversell how fast they can deploy the systems and how significant the business benefits will be. Overpromising and under delivering is the surest way to get the business to walk away from any technology, and it often sets back the use of the particular technology within an organization for a long time—even if many other enterprises are achieving success. In addition, when you set expectations that the benefits will come easily and quickly, business executives have a tendency to underestimate the required level of involvement and commitment. And when a sufficient resource commitment isn't there, the expected benefits usually don't come easily or quickly—and the project is labeled a failure.

BABY STEPS

It is said that every journey begins with the first step, and the journey toward creating an effective Big Data analytics holds true to that axiom. However, it takes more than one step to reach a destination of success. Organizations embarking on Big Data analytics programs require

a strong implementation plan to make sure that the analytics process works for them. Choosing the technology that will be used is only half the battle when preparing for a Big Data initiative. Once a company identifies the right database software and analytics tools and begins to put the technology infrastructure in place, it's ready to move to the next level and develop a real strategy for success. The importance of effective project management processes to creating a successful Big Data analytics program also cannot be overstated. The following tips offer advice on steps that businesses should take to help ensure a smooth deployment:

- **Decide what data to include and what to leave out**.

By their very nature, Big Data analytics projects involve large data sets. But that doesn't mean that all of a company's data sources, or all of the information within a relevant data source, will need to be analyzed. Organizations need to identify the strategic data that will lead to valuable analytical insights. For instance, what combination of information can pinpoint key customer-retention factors? Or what data are required to uncover hidden patterns in stock market transactions? Focusing on a project's business goals in the planning stages can help an organization home in on the exact analytics that are required, after which it can—and should—look at the data needed to meet those business goals. In some cases, this will indeed mean including everything. In other cases, though, it means using only a subset of the Big Data on hand.

- **Build effective business rules and then work through the complexity they create.**

Coping with complexity is the key aspect of most Big Data analytics initiatives. In order to get the right analytical outputs, it is essential to include business focus data owners in the process to make sure that all of the necessary business rules are identified in advance. Once the rules are documented, technical staffers can assess how much complexity they create and the work required to turn the data inputs into relevant and valuable findings. That leads into the next phase of the implementation.

- **Translate business rules into relevant analytics in a collaborative fashion**.

Business rules are just the first step in developing effective Big Data analytics applications. Next, IT or analytics professionals need to create the analytical queries and algorithms required to generate the desired outputs. But that shouldn't be done in a vacuum. The better and more accurate that queries are in the first place, the less redevelopment will be required. Many projects require continual reiterations because of a lack of communication between the project team and business departments. Ongoing communication and collaboration lead to a much smoother analytics development process.

- **Have a maintenance plan**.

A successful Big Data analytics initiative requires ongoing attention and updates in addition to the initial development work. Regular query maintenance and keeping on top of changes in business requirements are important, but they represent only one aspect of managing an analytics program. As data volumes continue to increase and business users become more familiar with the analytics process, more questions will inevitably arise. The analytics team must be able to keep up with the additional requests in a timely fashion. Also, one of the requirements when evaluating Big Data analytics hardware and software options is assessing their ability to support iterative development processes in dynamic business environments. An analytics system will retain its value over time if it can adapt to changing requirements.

- **Keep your users in mind—all of them**.

With interest growing in self-service BI capabilities, it shouldn't be shocking that a focus on end users is a key factor in Big Data analytics programs. Having a robust IT infrastructure that can handle large data sets and both structured and unstructured information is important, of course. But so is developing a system that is usable and easy to interact with, and doing so means taking the various needs of users into account. Different types of people— from senior executives to operational workers, business analysts, and statisticians—will be accessing Big Data analytics applications in one way or another, and their adoption of the tools will help to ensure overall project success. That requires different levels of interactivity that match user expectations and the amount of experience they have with analytics tools—for instance, building dashboards and data visualizations to present findings in an easy-to-understand way to business managers and workers who aren't inclined to run their own Big Data analytics queries. There's no one way to

ensure Big Data analytics success. But following a set of frameworks and best practices, including the tips outlined here, can help organizations to keep their Big Data initiatives on track. The technical details of a Big Data installation are quite intensive and need to be looked at and considered in an in-depth manner. That isn't enough, though: Both the technical aspects and the business factors must be taken into account to make sure that organizations get the desired outcomes from their Big Data analytics investments.