Balancing Innovation With Data Protection

The majority of my IT life has been focused on protecting and properly analyzing data. Some products and systems make this easier than others just as some people make this easier than others. When it comes to innovation, I have seen individuals try to use the need to move quickly as a justification to slack off on protecting data. This has never been my view because the hope and goal with innovation is to deliver solutions and services that, ultimately, are put to real use. Therefore, if data protection is not important during implementation, there is a strong probability that it will be missing when a solution coming from such effort goes to production. A good example is what we have seen with offerings dubbed the Internet of Things (IoT).[1] The focus there is usually on speed to market, not data protection.

Data at Rest/Data in Flight

Protecting data sounds easy, until you get down to the 1s and 0s of doing it. This is true whether we are looking at well-known, established products or cutting our teeth on the edge of innovation. The main difference between the two is that with established products, we typically can leverage security benchmarks, white papers, consulting firms with significant experience in securing those systems, and other resources to aid our efforts at protecting data. Innovation, by its very nature, is less likely to have those assets available.

The short answer to this dilemma is to design security in from the start. As we innovate, we should be asking the typical questions around data throughout the process:

- What data are stored?
- How are those data stored?
- Should any of the data be encrypted?
- How are the data encrypted?
- What data are transmitted?
- How are they transmitted?
- Are secure protocols used?

But I would also recommend that we go further and ask more questions about what we are building and working with as we innovate, including:

- What are the known vulnerabilities?
- How might those be exploited?
- What does the surface area look like?
- How might an attacker exploit it?
- Can we scan and test for vulnerabilities?
- What mitigation options do we have?

If these questions sound like what we ask for traditional projects and systems, you are correct. My experience with innovation is that because we are looking to move quickly, we may fail to do due diligence and ask these questions. At the end of the day, though, the outside world does not care if an organization has a data breach on an old, stodgy enterprise resource planning (ERP) system or loses data through something new and cool. Likewise, it does not matter to the victim or the investor if the data breach happened with an on-premises system or data stored in the cloud. A data breach is a data breach.

So why emphasize these questions for innovation efforts? Speed of delivery is a key aspect of any innovation. Can we get it there faster than anyone else or, at least, before our major competitors? In that push for speed, due diligence can be neglected. I say

K. BRIAN KELLEY | CISA, CDPSE, CSPO, MCSE, SECURITY+

Is an author and columnist focusing primarily on Microsoft SQL Server and Windows security. He currently serves as a data architect and an independent infrastructure/security architect concentrating on Active Directory, SQL Server and Windows Server. He has served in a myriad of other positions, including senior database administrator, data warehouse architect, web developer, incident response team lead and project manager. Kelley has spoken at 24 Hours of PASS, the PASS Data Summit, IT/Dev Connections, SQLConnections, the TechnoSecurity and Forensics Investigation Conference, the IT GRC Forum, and at various SQL Saturdays, Code Camps and user groups.

"can be," because it is dependent on the organization, possibly down to the team. Also, due diligence might be harder to perform in an innovation effort because time windows are often smaller. A team might have one week to do a penetration test of a relatively static product. But for the latest innovation effort, that same effort may only be given one day in the schedule. Prioritization based on risk becomes key for protecting the organization.

Protecting Against Bias

The topics of data and innovation naturally lead to a discussion of artificial intelligence (AI) and machine learning (ML). When we process data, ultimately, there is going to be some bias introduced. For instance, we have quite a few studies that compare election polls to election results to determine "What went wrong?" and "How wrong were we?" One study found an absolute election bias of 1.5 percentage points across a review of 4,221 polls for 608 US state-level presidential, senatorial and gubernatorial races during the period from 1998-2014.[2] So, despite our best efforts, for some of the most important results, we know that bias makes its way into the final result. Therefore, it is crucial is to minimize that bias.

One could assume that if we assign data analysis to AI and ML, that much bias could be avoided. Unfortunately, that is not the case. A search of ML and bias displays articles on five,[3] six, seven or eight types of biases, and perhaps more. The point is that there is bias even when we ask a computing resource to do the data analysis. The bias may even be within the sample itself.

In the 2004 US presidential election, for example, when early exit poll data were used to forecast winners, there were significant issues with the way the data were interpreted. One of those issues was the false assumption that voters would vote consistently the same throughout the day. If the data had been used to spot early trends, the data set was appropriate. However, since the data were used to forecast winners, the data did not include a proper sampling of voters to get accurate results.[4]

Even if we are trying to avoid bias, it may still creep into data sets. As an example, if we are trying to eliminate racial bias, we may eliminate racial identification from the data set to be analyzed, but there are proxies for race that might escape notice. One such proxy is postal codes,[5] because housing tends to be segregated, for example, in the United States. Thus, postal codes, for the most part, often suggest data about race. Ultimately, when we look at innovation and data, we want to make sure we are ethical in how we handle that data, and much of that effort should be around eliminating bias.[6]

> The outside world does not care if an organization has a data breach on an old, stodgy ERP system or loses data through something new and cool.

Pros and Cons of Pass-Through Audits for Controls

There is a great deal of innovation happening in the cloud, whether through the use of Software-as-a-Service (SaaS) offerings, with a vendor hosting the application for us, or Infrastructure as a Service (IaaS), which lets us create full virtual machines from pieces and parts, or Platform as a Service (PaaS), which lets us combine components into an overall solution. Much of AI and ML fall into the last of the three, PaaS. After all, some computer models require

a significant amount of hardware, and cloud providers are able to provide such computational resources more cost effectively than maintaining them on-premises due to economies of scale and the fact that they can schedule resource utilization nearly all the time, meaning they can effectively bill for use of those resources around the clock.

However, relying on cloud vendors often means relying on their security controls rather than our own. For instance, a cloud provider is responsible for physical security of the data centers and the servers and storage they offer to customers; users of that cloud provider's services are not. However, users have certain expectations of what is acceptable and what is not. Moreover, external auditors have expectations, too. A better way of thinking about pass-through audits is as "compliance inheritance." This is the term used by the Cloud Security Alliance (CSA).[7]

In other words, for certain aspects of an organization's compliance requirements, it can inherit the controls and the testing of said controls from the vendor. Most large cloud providers have resources dedicated to reporting what standards, laws and regulations require their compliance. Customers can request the appropriate reports derived from controls testing and detailed breakdowns of what services and offerings meet which standards.

At first glance, all of that sounds great. In fact, it often is. Because of economies of scale, cloud providers are able to get more for their spend, resulting in greater security measures and offerings than what most organizations can provide on their own. Also, the local organization's politics do not apply, as the controls are being handled by a third party. From this perspective, pass-through auditing is a huge benefit to organizations.

However, the problem with pass-through auditing is that things are sometimes taken too far. Every major cloud provider has some form of shared responsibility model that indicates what the provider is responsible for and what the customer must handle. We have seen enough reported cases over the years to make it clear that organizations often selectively ignore that "shared" aspect of responsibility and pay a dear price for it. The cloud provider does not do it all.

Also, since it is a shared model, there are times when it may be difficult to determine if the provider's controls are enough or if the organization needs to add its own controls—that is, if the organization is allowed to do so. There are many things organizations do from a compliance perspective for their own resources beyond just physical security that are not allowed in the cloud. The key is to understand what each side is responsible for and work to clarify areas where things are not so clear. At the end of the day, the organization is still responsible for its data, regardless of where they are held.

We do not dodge bias issues just because a computing resource is doing the analysis.

Protecting Data Is Hard, But Necessary

With innovation, we understand that moving quickly is a necessity. However, we cannot neglect data protection. If anything, we need to include proper protection from the start. Innovation does not give license to bypass the usual questions that are asked when we have more time. It requires us to answer or evaluate them more quickly.

Innovation also means using new tools to analyze data in ways we have not before. AI and ML are great examples of "new" technologies that we now have at our disposal. However, we must remember that it is not enough to protect data. We must also use data responsibly and ethically. One of the greatest problems we face with data is bias. Some bias is easy to spot, such as sampling incorrectly. Other biases might be more difficult to identify and mitigate, such as the use of postal codes as a proxy for race. We do not dodge bias issues just because a computing resource is doing the analysis.

Finally, more and more innovation is headed to the cloud. As a result, we must scrutinize what cloud providers offer in the way of controls. We do not simply look, we have to rely on them, too, in what we call pass-through audits. Pass-through audits are tricky because the cloud provider does not

LOOKING FOR MORE?

- Read *Defending Data Smartly. www.isaca.org/defending-data-smartly*
- Learn more about, discuss and collaborate on governance in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*

handle everything. The overall solution requires a shared responsibility. It is important to understand that shared responsibility and know what the cloud provider is responsible for and what belongs to our organization. Sometimes this gets murky, but we still have to work it out. After all, if there is a data breach, it is the organization that will make headline news.

Endnotes

1 Reed, J.; "IoT Security and the Internet of Forgotten Things," *Security Intelligence*, 22 March 2022. *https://securityintelligence.com/articles/iot-security-internet-forgotten-thing/* .

2 Houshmand, S.-M.; D. Rothschild; S.Goel; A. Gelman; "Disentangling Bias and Variance in Election Polls," *Journal of the American Statistical Association*, vol. 113, no. 522, 2018, p. 604–614, *https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1448823*

3 Metwalli, S. A; "Five Types of Machine Learning Bias Every Data Scientist Should Know," *Towards Data Science*, 24 February 2021, *https://towardsdatascience.com/5-types-of-machine-learning-bias-every-data-science-should-know-efab28041d3f*

4 Anderson, N.; F. Fiore; "Early Data for Kerry Proved Misleading," *Los Angeles Times*, 4 November 2004, *https://www.latimes.com/archives/la-xpm-2004-nov-04-na-pollsters4-story.html*

5 George, A.; "Thwarting Bias in AI Systems," Carnegie Mellon University College of Engineering, Pittsburgh, Pennsylvania, USA, December 2018, *https://engineering.cmu.edu/news-events/news/2018/12/11-datta-proxies.html*

6 Chugh, V.; "Handling Data Bias: A Journey Towards Ethical AI," *Towards Data Science*, 5 July 2022, *https://towardsdatascience.com/handling-data-bias-9775d07991d4*

7 Mogull, R. *et.al.*; *Security Guidance: For Critical Areas of Focus In Cloud Computing 4.0.* Cloud Security Alliance, USA, 2017, p. 56, *https://cloudsecurityalliance.org/download/security-guidance-v4/*