CAPITAL ONE UNIVERSAL SEARCH

DATA UNIFICATION TO BENEFIT CYBERSECURITY DEFENSE AND COMPLIANCE

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Andrew Li

December 9, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Rider Foley, Department of Engineering and Society

Rosanne Vrugtman, Department of Computer Science

Introduction

Detailed and comprehensive data collection has become pervasive in nearly every sector imaginable. Storing and retrieving useful insight from this data is the challenge of “big data.” A central task in maintaining such sheer quantity and diversity of data is organizing and unifying it. Data fragmentation and lack of unification can limit usability of data and cause severe security vulnerabilities (Gibbs et al., 2002). This issue is particularly important in environments where security and safety are critical, such as when dealing with financial or health data. When data is stored in very disparate systems, called data silos, and protected under inconsistent security standards, information that can be used to rapidly detect security breaches is inefficiently used (ThreatConnect, n.d.). Sensitive data also becomes vulnerable to unauthorized access from multiple, insecure endpoints. Both issues can be alleviated by data unification.

Data unification refers to the consolidation of different types of data often stored in separate virtual or physical databases into a single, accessible source. As data collection and storage systems have matured in recent years, data unification has become both increasingly difficult and more important to enforce—systems that were designed incrementally and independently become resistant to unification. Redesigning or merging these systems can be expensive and complex—but maintaining fragmented databases is very expensive in itself (Mallikarjuna, 2020). Financial companies in particular have struggled with the costs of data fragmentation especially as they have become popular targets for cyber attacks. Cyber attacks can be extremely costly with legal penalties and damage to reputation, both costs that are exceptionally high for financial companies (Mallikarjuna, 2020).

Data unification becomes even more complicated with the movement of data to cloud storage and computing providers like Amazon Web Services (AWS) and Microsoft Azure. With the addition of another third-party entity involved in the storage and access of data, more endpoints vulnerable to attack are opened. The increasing prevalence of big data in tandem with cloud computing introduces a worsening data management environment for cybersecurity (Campos, 2016). Along with this trend has come a string of high-profile and extremely costly data breaches at financial institutions in the past few years.

An instructive example is that of Capital One's data breach in 2019, which occurred during March but was not detected until July 19 (Capital One, 2022). The case of Capital One's breach touches on many of the broader patterns in data management and the problems that they pose. Since 2014, Capital One has been prioritizing a "serverless" architecture for their services by moving all data and computation to the Amazon Web Services cloud platform (Novaes Neto et al., 2020). Capital One completed this migration in 2020, closing its last data center and becoming the first US bank to move completely to the cloud (Amazon Web Services, 2020). This milestone was achieved nearly a year after the data breach in 2019, which occurred when an ex-AWS employee gained unauthorized access to Capital One and many other institutions' data hosted by AWS by exploiting a firewall misconfiguration (Novaes Neto et al., 2020). Some estimates place the total cost of the breach to Capital One at $500 million (Lu, 2019). The data breach was a result of increased attack space due to Capital One's reliance on cloud services, inconsistent security standards implemented across data stores, and insufficient monitoring and analysis of network and database sensor data. Data unification is critical in solving all these factors and will need to be a priority for companies and institutions to prevent data breaches in the future.

Cybersecurity Data Unification

Like many companies, Capital One employs a dedicated cybersecurity team to prevent and respond to cyber attacks. Cyber teams rely on data collected from a diverse array of sensors and sources including both automated software tools and manual generation by associates. Automated software includes built-in monitoring tools that record and analyze access to the company's networks and logins to services both internal and third-party. This automatic collection of data is necessary for both algorithmic detection of insecure and malicious activity as well as detection by analysts in a cybersecurity operations center (Vieira et al., 2020). This data is the frontline defense for companies, as it can potentially allow for rapid detection and response to incoming threats. The data can also be a liability, as security-critical information collected internally can also be a prime target for attackers (Rawat et al., 2021).

Non-sensitive cybersecurity information like news reports of cyberattacks or newly established cybersecurity guidelines are also critical for the overall security of a company. This information needs to be organized and easily accessible for all employees, not just those that work directly with critical data and systems. Human-error, especially by those not formally trained in cybersecurity, is a significant cause of data breaches and can be mitigated by increasing awareness and knowledge of the general cybersecurity space (Coffey, 2017).

At Capital One, both sensitive and non-sensitive cybersecurity information can be accessed by associates using internal websites. One website may provide access to both types of data, prompting the need for strict access controls to ensure that only authorized users may view more sensitive information. For internal cybersecurity analysts and other privileged users, efficiently accessing and filtering information on these sites can be an important element of their everyday job. For employees in unrelated fields, being able to find information quickly and conveniently is critical to promote general awareness about cybersecurity developments and practices. However, many of these sites do not provide that level of power and ease of use, especially when the site must coalesce data of many different sources, types, and access controls. In one case, data of different types were isolated to separate pages on the site, with no way to search for general keywords or subjects across types. For instance, it was not possible to retrieve information about both threat actors from China at the same time as cyber incidents involving China. While at Capital One, I worked to unify disparate types of data on one of these internal websites into a single endpoint that could be used by both associates in all levels and roles. This project included a powerful search engine that could access every data record stored by the site and unify it into a single view where results could be sorted and filtered. By implementing strict access controls, the feature could be a central repository for associates across the entire company. Unifying the data on the site would increase the productivity of associates who use the site, potentially allowing them to better detect and analyze security risks and generally increasing security literacy across the organization. This will allow data to be used more proactively and defensively in preventing cyber attacks (Vieira et al., 2020). This project can also aid in supporting Capital One's compliance with cybersecurity reporting requirements that have been established and strengthened since its 2019 breach (Gibbs et al., 2002; Novaes et al., 2020).

At Capital One and other companies, various entities including people, systems, and organizations are involved in the production and consumption of just as many different types of data. Innovation and change are needed to improve this complex relationship, but change can also be costly and confusing. The success of the feature I developed also depends on its ability to harmonize the different relationships between each of those entities without disrupting the established flow of data between them.

Actor Network Theory in Data Creation, Consumption, and Management

Actor Network Theory, as proposed by Bruno Latour (1992), can be used to understand the social groups and technologies involved in data unification. Actor Network Theory argues that technology is shaped through the interactions between people, institutions, and the technology itself. In the case of data unification, these actors may generate data that others consume as well as consume data that other actors generate, creating a complex network where data of different types and uses flow between nodes. Between the technical and non-technical actors, tasks and responsibilities can be delegated to technology or prescribed back to the humans, creating a web of information flow between people and computers. Within an institution or company, data may be collected, stored, and retrieved by both human associates and computers. Cybersecurity analysts monitor incoming data regarding cybersecurity-critical actions like logins and access tokens and analyze this information to create new data like reports and alerts. Non-cyber associates may search for and retrieve information like internal risk reports, news articles, and blog posts. Automated systems process data to detect and remedy abnormalities and potential cyber attacks. Customers and users generate sensitive data and personal information that is stored by financial institutions. As data storage and computing services move to third-party cloud providers, these providers also enter the network of data stored and used by financial institutions. Finally, governments and agencies regulate the ways companies and employees can access and store data as well as the technical systems themselves. The ways in which these actors interact with each other are mediated by the successes and failures of existing technology.

The main goal of data unification is to increase the efficiency of processing and analyzing the data. Thanks to the rapidly increasing sophistication and scale of computing technology, it is no longer feasible for humans to sort and organize massive amounts of data. Data unification involves creating technical systems including databases and their supporting tools that can understand data at a high enough level to link together different types of data and structure them into a single conceptual and virtual database. Thus, data unification aims to delegate the responsibility of organizing and maintaining data to technical systems, rather than human cybersecurity analysts or staff more generally (Stonebraker, 2017). In response, technology prescribes certain responsibilities back to humans. For instance, data that is manually entered into a data store, such as cybersecurity event alerts submitted by cyber analysts, must be formatted correctly with certain fields present to be processed by the technical system. If data doesn't comply with the specifications set by the technical system's understanding of different types of data, the system will be unable to process it and enter it into the database. Depending on the complexity of this data handoff between human and computer, the technology may discriminate based on the user's familiarity with computer systems and programming. A well-designed and accessible system for entering data can be used by any employee in a company, whereas more arcane systems may require knowledge of data formatting languages like JSON or XML. The balance between what is delegated to technology and what is prescribed to humans determines this discrimination, but such a compromise is unavoidable. Thus, in the process of creating a structured and orderly database, some structure must be created by humans, and some by technology.

The idea of delegation of responsibilities is a very useful framework for examining the relationships between other actors as well. The migration of internal data centers to cloud providers, like that completed by Capital One, involves a delegation of physical systems to companies that specialize in data storage. In response, the cloud provider is responsible for creating secure endpoints where companies like Capital One can develop tools and access data. This delegation is a result of industry specialization to increase efficiency and ease of use (Tak et al., 2011). It also represents some level of data unification under more consistent storage practices by a third-party, but creates corresponding security risks as well. The movement to cloud providers has spawned a lucrative industry—with a market size of $368.97 billion in 2021 (Grand View Research, 2021)—for big, centralized providers like Amazon Web Services and Microsoft Azure. Companies like Capital One, in contrast, have shuttered their data centers.

The relationship between companies, consumers, and governments can also be characterized by delegation and prescription. Consumers are expected to assign away certain rights to their data when agreeing to privacy policies and benefiting from the services provided by companies (Palmieri III, 2019). Whether or not potential users agree to the policy leads to an implicit negotiation between the benefits of the company's services and the costs of lost privacy and potential data leaks. Whether or not companies are liable for misuse of that data, and to what degree, is highly variable depending on the jurisdiction. The policies of the United States, European Union, and China all vary significantly in their level of responsibility delegated to the company (Palmieri III, 2019). Different methods of regulation have different advantages and disadvantages which can depend on the social norms established by law and by custom (Martínez-Martínez, 2020). Private companies may resist the regulatory relationship with governments to reduce their liability for cyber attacks and data breaches (Peng, 2018), creating a potentially adversarial relationship between these two actors.

Technology in data unification must establish its own delegated and prescribed responsibilities as well as manage the relationships between other actors. For instance, government regulation of private companies often includes data reporting requirements that depend on the data collection and organization of the technical system. Data unification can define and change the relationships between technology, people, and institutions.

Research Question and Methods

How has lack of data unification in financial institutions and companies influenced their vulnerability to cyber attacks and compliance with cybersecurity regulation? This question is critically important for organizations of all kinds facing increased aggression and crime in cyberspace. Data is currency, and financial and health data are gold mines for cyber criminals in an underground market (Wirth, 2017). As a result, financial institutions and companies are a major target for cyberattacks—and are liable to lose substantial amounts of money if their data is improperly protected. An increasingly global and complex Data unification is a crucial element of successful cyber defense as data proliferates. To investigate the current state of data unification in cybersecurity, its impact on past cyber breaches, and identify areas of potential improvement, I will use a combination of interviews and case studies in my paper. Specifically, I will interview staff in the University of Virginia's (UVA) Information Security office regarding UVA's past security incidents and current cybersecurity strategies. To the extent possible, I will aim to understand the types of data UVA collects on its students, faculty, and staff, how that data is stored, how it is accessed, how it is protected, and which laws regulate these processes. Additionally, I will inquire about UVA's history with data breaches, including the 2015 hack allegedly by Chinese attackers, 2016 phishing breach, and 2018 health system data breach.

In addition to interviews, I will conduct case studies on past cyberattacks on financial institutions and companies including Capital One in 2019 and Equifax in 2017. These high-profile cases have been studied heavily and have more publicly available information about them. These case studies will focus on the organization that suffered from the hack rather than the attackers and investigate which data management weaknesses allowed the attack to succeed and what actions were taken by the organization afterwards to prevent similar breaches in the future. Additionally, the actions and reactions by regulators before and after the breaches will be analyzed to understand the role of policy in shaping data unification. Case studies are a particularly effective research method for contemporary events to answer the critical questions of how and why these events happened (Yin, 2009).

Conclusion

Today's cybersecurity environment poses significant challenges for a data-driven world. Data unification—or lack thereof—by organizations is becoming increasingly visible during failures. The technical project provides an example of how data unification can be achieved during data retrieval, even when the underlying data may be fragmented. Further improvements to this project will improve user accessibility, particularly for associates who do not have a technical background. The paper will investigate the broader trends and implications of data unification within cybersecurity, drawing on specific cases and primary sources to illustrate the growing importance of data unification for cyber defense and regulatory compliance. The paper will reinforce the need for synergy between companies, consumers, governments, and technology to solve this problem.

References

2019 Capital One Cyber Incident | Frequently Asked Questions. (2022, April 22). Retrieved October 26, 2022, from Capital One website: https://www.capitalone.com/digital/facts2019/faq/

Campos, J., Sharma, P., Jantunen, E., Baglee, D., & Fumagalli, L. (2016). The Challenges of Cybersecurity Frameworks to Protect Data Required for the Development of Advanced Maintenance. *Procedia CIRP*, *47*, 222–227. https://doi.org/10.1016/j.procir.2016.03.059

Capital One Completes Migration from Data Centers to AWS, Becomes First US Bank to Announce Going All In on the Cloud. (2020). Retrieved October 26, 2022, from Amazon Web Services website: https://aws.amazon.com/solutions/case-studies/capital-one-all-in-on-aws/

Coffey, J. W. (2017). Ameliorating Sources of Human Error in CyberSecurity: Technological and Human-Centered Approaches. *The 8th International Multi-Conference on Complexity, Informatics and Cybernetics*. Presented at the IMCIC, Pensacola.

Gibbs, M. R., Shanks, G., & Lederman, R. (2002). Data Quality, Database Fragmentation and Information Privacy. *Surveillance & Society*, *3*(1). https://doi.org/10.24908/ss.v3i1.3319

Grand View Research. (2021). *Cloud Computing Market Size, Share & Trends Analysis Report By Service (IaaS, PaaS, SaaS), By Deployment (Public, Private, Hybrid), By Enterprise Size, By End Use (BFSI, IT & Telecom, Retail & Consumer Goods), By Region, And Segment Forecasts, 2022 - 2030* (p. 130). San Francisco, California.

Latour, B. (1992). Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts. In *Inside Technology*. *Shaping Technology / Building Society: Studies in Sociotechnical Change*. Cambridge, Massachusetts: MIT Press.

Lu, J. (2019, August 15). *Assessing The Cost, Legal Fallout Of Capital One Data Breach* [SSRN Scholarly Paper]. Rochester, NY. https://doi.org/10.2139/ssrn.3438816

Mallikarjuna, J. (2020). *The Relevance of Data Unification*. SG Analytics.

Martínez-Martínez, D.-F. (2018). Unification of personal data protection in the European Union: Challenges and implications. *El Profesional de La Información*, *27*(1), 185. https://doi.org/10.3145/epi.2018.ene.17

Novaes Neto, N., Madnick, S., Moraes G. de Paula, A., & Malara Borges, N. (2020, March 1). *A Case Study of the Capital One Data Breach* [SSRN Scholarly Paper]. Rochester, NY. https://doi.org/10.2139/ssrn.3570138

Palmieri III, N. F. (2019). Data Protection in an Increasingly Globalized World. *Indiana Law Journal*, *94*(1), 297–329. Retrieved from https://www.repository.law.indiana.edu/ilj/vol94/iss1/7

Peng, S. (2018). “Private” Cybersecurity Standards? Cyberspace Governance, Multistakeholderism, and the (Ir)relevance of the TBT Regime. *Cornell International Law Journal*, *51*(2), 446–470. Retrieved from https://scholarship.law.cornell.edu/cilj/vol51/iss2/4

Rawat, D. B., Doku, R., & Garuba, M. (2021). Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. *IEEE Transactions on Services Computing*, *14*(6), 2055–2072. https://doi.org/10.1109/TSC.2019.2907247

Stonebraker, M. (2017). *The Seven Tenets of Scalable Data Unification*. Cambridge, MA: Tamr.

Tak, B. C., Urgaonkar, B., & Sivasubramaniam, A. (2011). To Move or Not to Move: The Economics of Cloud Computing. *3rd USENIX Workshop on Hot Topics in Cloud Computing*. Presented at the USENIX.

ThreatConnect. (n.d.). *Fragmentation: The "Silent Killer" of Your Security Management Program*. Retrieved from https://threatconnect.com/wp-content/uploads/ThreatConnect-whitepaper-fragmentation.pdf

Vieira, K., Koch, F. L., Sobral, J. B. M., Westphall, C. B., & Leao, J. L. de S. (2020). Autonomic Intrusion Detection and Response Using Big Data. *IEEE Systems Journal*, *14*(2), 1984–1991. https://doi.org/10.1109/JSYST.2019.2945555

Wirth, A. (2017). The Economics of Cybersecurity. *Biomedical Instrumentation & Technology*, *51*(s6), 52–59. https://doi.org/10.2345/0899-8205-51.s6.52

Yin, R. K. (2009). *Case Study Research: Design and Methods* (Fourth edition).