RAPID7

RESEARCH

2021 Cloud Misconfigurations Report

Rapid7 Research

September, 2021

TABLE OF CONTENTS

Executive Summary

In the middle of 2021, as the working world for knowledge workers seems to be returning to normal, Rapid7 researchers took a look at 121 publicly reported data exposure incidents that were disclosed in 2020 to see if we could find some common causes and circumstances among them.

As a result of this research, we found that:

- Of the 121 published incidents, 15 industries were represented among the affected organizations, with Information, Entertainment, Professional, and (perhaps most worryingly) Healthcare being the most represented in the data set.

- An array of 14 information types (including "other") were reported exposed; most notably, datasets concerning credentials (usernames and passwords), personal financial information, and personal health information were among the reported incidents.

- Through 2020, we saw an average of about 10 incidents a month reported, and a preponderance of these incidents (62%) were discovered by independent researchers (rather than criminal attackers). Notably, 35% of the total incidents were sourced from only two specific individual researchers.

- The most common type of exposure reported was insufficiently protected Amazon Simple Storage Service (S3) buckets and Elasticsearch databases, which account for 45% of all reported exposures in 2020.

For more detail on the corpus of reported events, please see **Appendix A**.

Introduction

Despite the promises of power and productivity, moving business processes and mission-critical data to the cloud can be perilous if one overlooks key safety and resilience configurations and controls. When those misconfiguration missteps occur, data can be exposed, leaving organizations with the unpleasant tasks of breach response, not to mention the regulatory and legal consequences if that data is personally identifiable information, sensitive health information, or other specially sensitive categories of data.

Our 2021 Cloud Misconfigurations Report documents the commonalities and patterns associated with leaks and breaches disclosed during 2020 in order to help organizations understand common misconfigurations, ultimately so organizations can avoid making the same missteps as they embark or continue on their cloudy journeys.

Rapid7 researchers identified 121 publicly reported data exposure incidents that were disclosed throughout 2020. On average, there were 10 disclosed incidents per month across 15 industries.[1]

Most reported incidents were discovered, disclosed, and remediated within the same month (often within the same week) primarily due to the fact that they were actively discovered by individuals looking for poorly secured services. Some took years to discover and disclose, mostly due to the nature of the breach method (one was esoteric based on how analytics services were configured, and others had no details of breach origin, just how old the data was and when it showed up on criminal forums). This may chafe against our collective intuition, but it tracks with statistics from the past few Verizon Data Breach Investigations Reports, which continue to indicate an upward trend in defenders detecting evidence of breaches at a faster rate for certain types of incidents.

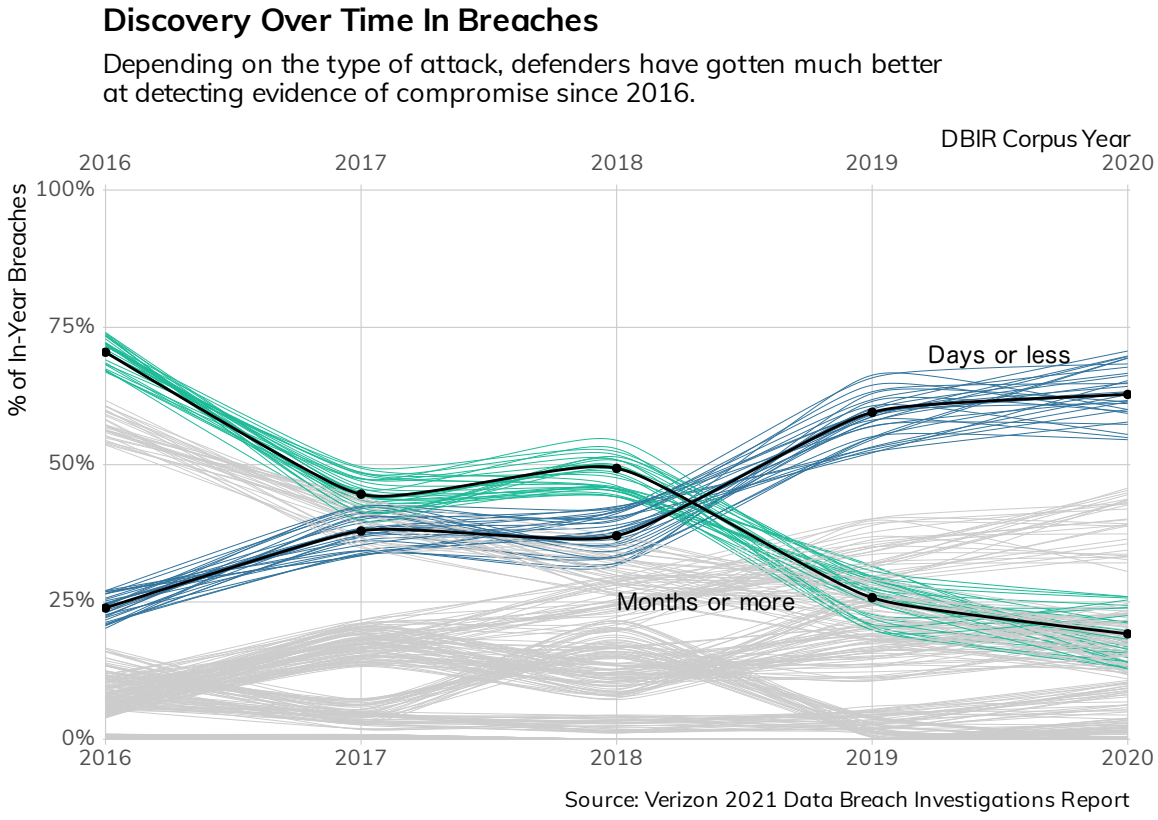

Figure 1: Discovery over time for breaches from the Verizon Data Breach Investigations Report Corpus

[1] We use the North American Industry Classification System (NACIS) when assigning industry names to organizations. (http://www.census.gov/eos/www/naics/2012NAICS/2012_Definition_File.pdf)

While the disclosed medium was unspecified in nearly half of the incidents, misconfigured permissions on AWS "buckets" and internet-facing Elasticsearch servers accounted for 25% and 21% (respectively) of reported incidents. This isn't surprising, given how easy it is for any individual to discover these services via easily accessible services such as GrayHat Warfare[2] and Shodan.[3]

The median data exposure was 10 million records, though one "mega breach" resulted in the exposure of over 20 billion-with-a-B records.

To avoid suffering an incident from similar cloud misconfigurations, companies should prioritize the adoption of a new model of security that provides continuous enforcement of controls and ensures secure configurations of all cloud services. Note that this activity is not a "set it and forget it" task, and all current and new cloud resources should be monitored and have policies enforced continually to avoid even a temporary exposure of these often dynamic environments.

Finally, we cannot help but notice that 75 of these incidents (62%) were initiated by security researchers trawling the internet for unprotected cloud S3 buckets and unsecured databases, and two people in particular make up 35% of these researcher-initiated disclosures.

As a quick aside, while not every organization has a vulnerability disclosure process (VDP), Rapid7 researchers have found that making a sincere effort to disclose issues like these privately first can be an excellent opportunity to help organizations kickstart their VDP and be more responsive to future issues.

In other words, we believe that citizen researchers also seeking to reduce sensitive data exposure on the internet should take a sympathetic stance when it comes to disclosure — as the individuals in this corpus have done — and give organizations an opportunity to do the right thing before a public disclosure.

[2] GrayHat Warfare https://buckets.grayhatwarfare.com/
[3] Shodan https://shodan.io/

Attack and Disclosure Cadence

Internet-facing service misconfigurations have been happening since the first devices were wired up to ARPANET[4], but the mass adoption of cloud computing has significantly increased the regularity of these missteps, especially in the early years of "the cloud," where the default configurations for many turnkey services, such as databases and file storage, were quite permissive. Despite the adoption of safer default service configurations in modern cloud computing environments — such as AWS, Google Cloud, and Microsoft Azure — misconfigurations still occur, and do so with distressing frequency.

While some misconfigurations are due to human error or caused by the assumption that software components come with safe defaults, many are deliberate choices to make it easier to access a given resource. Criminal attackers and opportunistic researchers alike know this all too well, which helps explain the frequency of monthly disclosures:

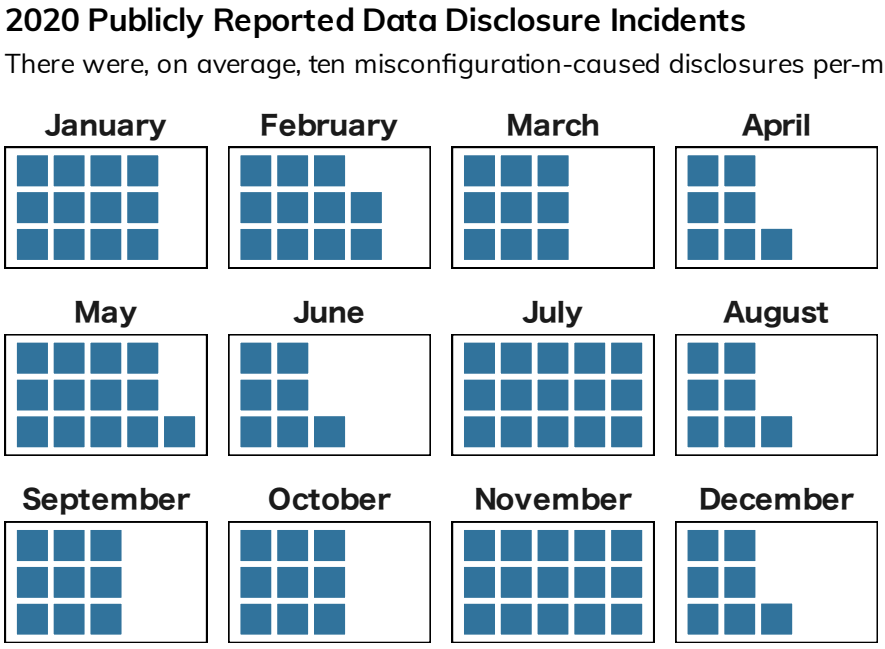

Figure 2: 2020 Monthly Cloud Misconfiguration Data Disclosure Incidents

The fact that these discoveries occur so frequently should alone be sufficient evidence to convince organizations and individuals to ensure the safety and resilience of their internet-facing services and data. However, many of the reports are only posted on niche outlets and rarely rise to the level of broader regional or national news.

[4] The first ARPANET outage was due to a misconfigured network service, documented in RFC 789: http://www.faqs.org/rfcs/rfc789.html

Scratched Records

Each incident in our 2020 corpus saw records exposed to the prying eyes of criminal actors or "researchers." Virtually no industry was immune to attack, and no record type was safe from disclosure.

The figure below shows the breakdown of record types disclosed by industry. Information, Entertainment, Professional, and Healthcare organizations had the widest array of record types exposed, but all industries exposed records of more than one type.[5]

Record Loss Type by Industry

Each cell represents the number of organizations exposing a given record type in an incident.

Information, Entertainment, Professional, and Healthcare organizations had the widest array of record types exposed but all industries exposed records of more than one type.

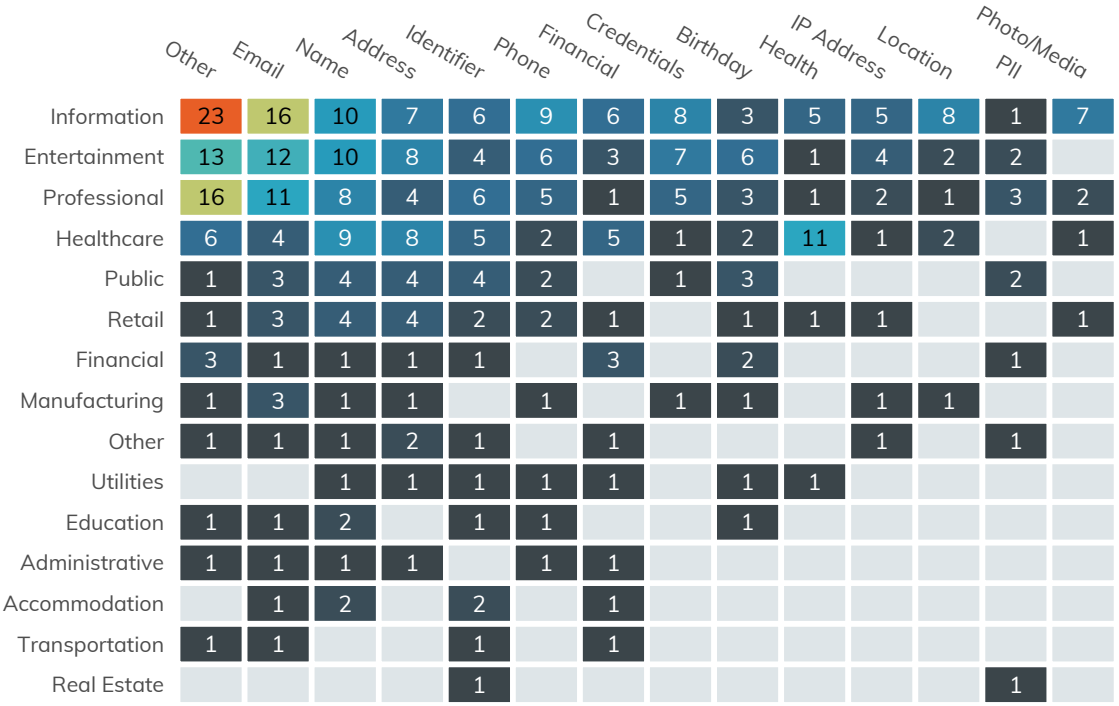

Figure 3: 2020 Cloud Misconfiguration Record Type Disclosures by Industry

The above chart also reveals that we, as individuals, seem to be okay with handing over a wide array of details about ourselves to many organizations. Given the frequency of record disclosure, perhaps we should consider being more judicious about what we share and with whom we share it.

[5] The "Identifier" type covers anything that could uniquely identify an individual out of a group, such as a driver's license number, bank account number, etc.

When records are disclosed, they tend to be lost en masse, with distribution numbers that would be the envy of even the highest grossing pop artist.[6] While less than half of the misconfiguration disclosures in 2020 would make it to the RIAA charts, 52 incidents snagged gold, platinum, or diamond honors.

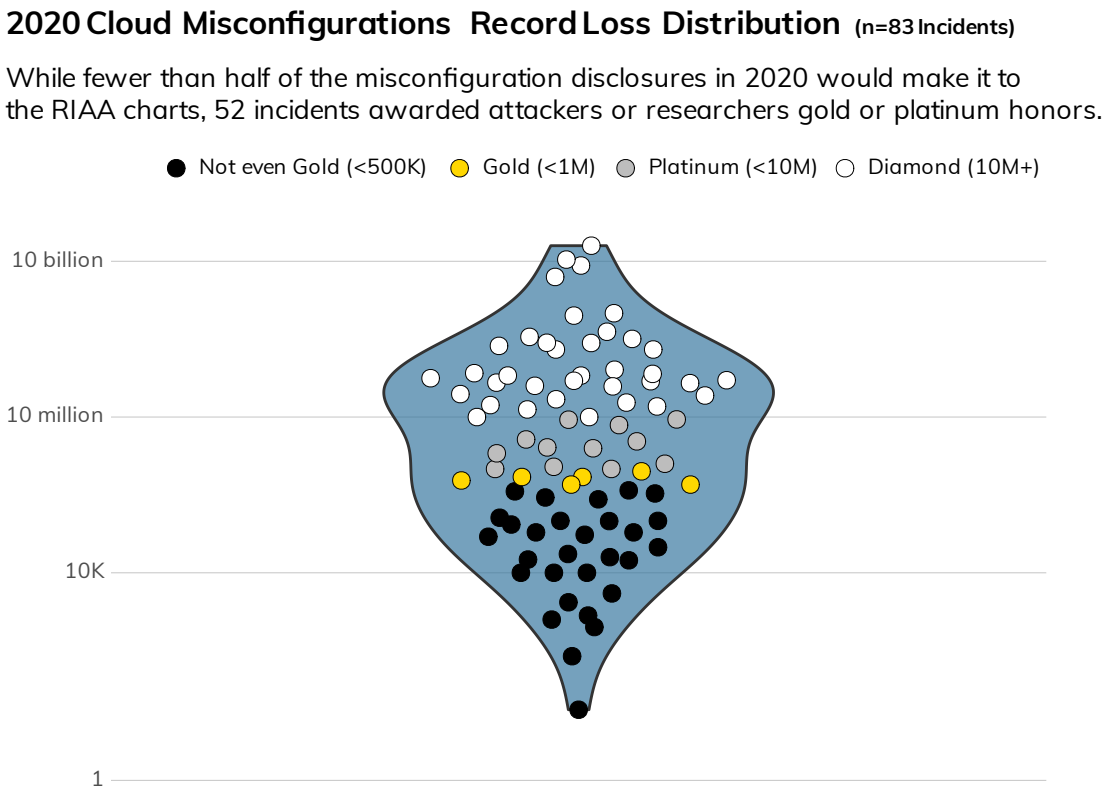

Figure 4: 2020 Cloud Misconfiguration Record Disclosures

Though the regular cadence of massive breaches has given us all a case of "breach fatigue," we should not treat these loss events so casually. While attackers may be making serious coin lately from ransomware attacks [7], our personal, financial, and health data are still valuable commodities that many organizations appear to be treating quite casually. It is vital that we all hold every business, agency, and non-profit accountable for ensuring the safety of our data in their cloud environments.

[6] It's Taylor Swift, by the way. Also known as @SwiftOnSecurity. See <https://www.statista.com/statistics/282151/highest-paid-musicians/> and <https://twitter.com/SwiftOnSecurity>

[7] https://www.bloomberg.com/news/articles/2021-05-20/cna-financial-paid-40-million-in-ransom-after-march-cyberattack

Easy Discovery and Quick Fixes

As previously noted, there are products and public services that make it trivial to discover exposed cloud resources. This ease of discovery usually leads to quick disclosure and remediation — apart from breach response and notifications — since the exposure was likely caused by a single errant configuration setting, as shown in the figure below:

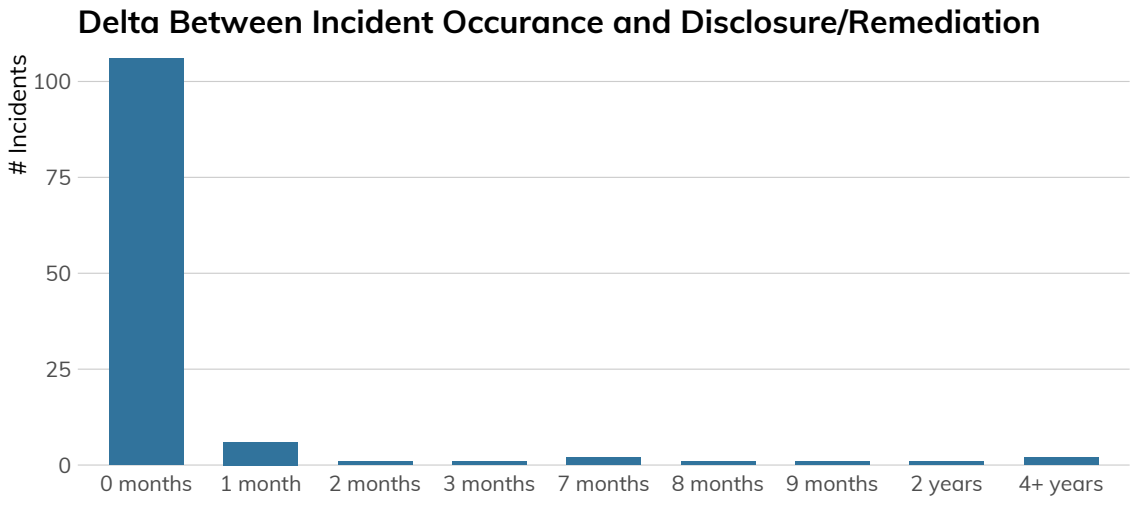

Figure 5: Disclosure/Remediation Distribution

Incidents with longer disclosure/remediation time-frames tend to be caused by criminal or malicious actors, while the incidents with shorter turn-around times tend to be reported by opportunistic researchers. In all cases, the misconfigured resource was in said errant state for the indicated time-period and only fixed after disclosure, placing further emphasis on the need for continuous monitoring and settings compliance.

Speaking of services, AWS S3 file/object buckets and Elasticsearch databases were still the favorites of criminal attackers and opportunistic researchers alike, accounting for nearly 45% of the misconfigured and compromised technologies:

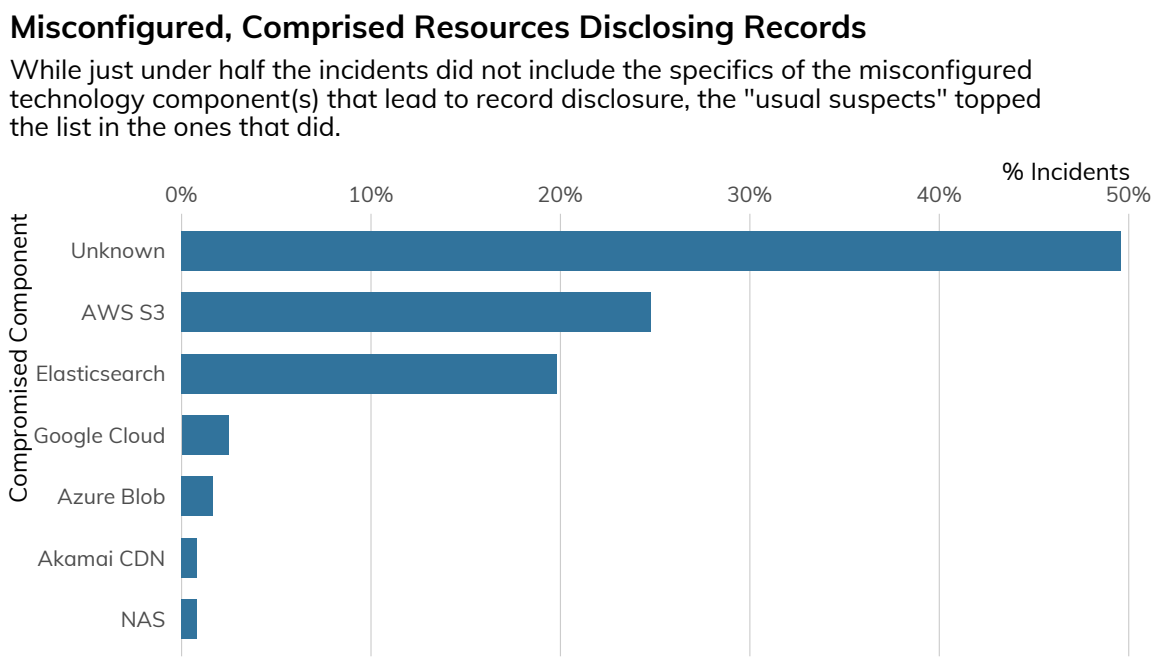

Figure 6: Misconfigured, Comprised Resources Disclosing Records

Both components now come with sane and safe defaults, which means the misconfiguration was very likely deliberate or at least set without fully understanding the risk. This further shows that the continuous monitoring and compliance enforcement of cloud resources should be configured to check for and set safe and resilient settings, not just the ones you believe to be correct.

Industry Breakdown

Finally, the chart below outlines the collected observations by industry — Information, Professional, Healthcare, and Entertainment top the list of affected industries, with a significant drop-off for other NACIS-defined industries. We're not sure why these industries seem to be more at-risk to these misconfigurations, but knowing that they are, organizations in these industries should take special care to make sure their configuration settings are sane and safe.

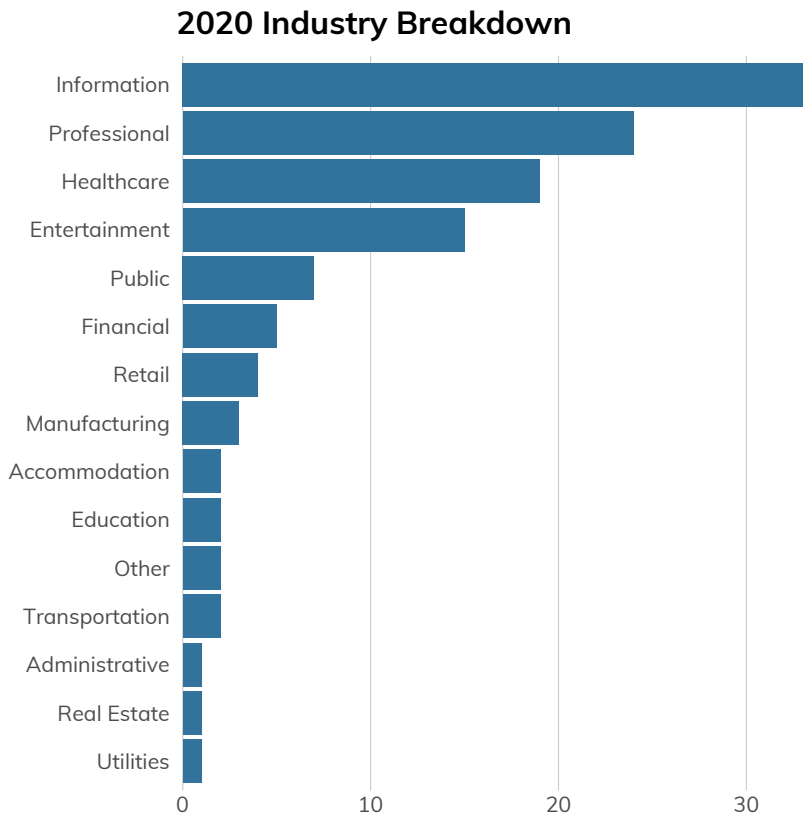

Figure 7: Industry breakdown affected by publically reported cloud misconfiguration breaches

Conclusion

Way back in 1711, Alexander Pope, the now infamous poet and satirist, coined the phrase “To err is Human;”, and the findings in this report provide plenty of evidence that we humans continue to err just fine these 310 years later, bringing our missteps all the way with us to the cloud. While we can all be forgiven for making mistakes, there are some concrete steps we can and should take to prevent our own organizations from garnering media attention and ending up on next year’s misconfigurations list.

First and foremost, you should now be keenly aware that there are individuals actively seeking out cloud service misconfigurations on a daily basis. Given the right tooling, it’s almost trivial for any moderately clever person to hunt for these cracks in the cloud at scale, and they don’t even need to be targeting your organization specifically to come across that unintended misconfiguration which ends up exposing sensitive data in your care.

The good news is that with some planning — i.e. you know what you’re exposing and also know what safe and resilient configurations should be in place — and automation — i.e. you have automated processes in place to monitor

said configurations and both alert on and remediate the errors when found — you can avoid seeing your organization's name in breach headlines.

Second, you should now also see that no industry or organization is immune. Misconfiguration-caused breaches in 2020 happened in every industry and across organizations of all shapes and sizes. Your ten-person startup can make the same mistakes as a 100+ year old brand, but you both have access to straightforward solutions that can help you avoid missteps due to misconfigurations.

Finally, your past mistakes may come back to haunt you if malicious attackers happened to see that errant configuration you left exposed for just a day four years ago. While you do need to shore up your present-day cloud-based services, you should ensure there is an entry in your incident response tabletop exercise playbook for practicing how you would respond to a modern-day dump of a years-old incident.

Taking the weight of worrying about misconfigurations off your mind will help free your organization to deliver solutions and innovate (securely) at scale.

Appendix A: Source Material

All of the catalogued instances are available for public inspection at <https://github.com/rapid7/data/tree/master/2021-cloud-misconfigurations>. Since this is a public GitHub repository, we would love to know if we missed any public reporting. Specifically, our corpus here seems particularly United States-focused, and we'd love to know if we missed any major incidents that might have been reported in other countries and other languages.

Furthermore, if you find this kind of research useful, we invite you to join our effort in collecting these stories periodically — if you're interested, just follow the Rapid7 Labs data repository at <https://github.com/rapid7/data>, and toward the end of the year, we expect to open up collection for 2021 reports of cloud-based exposures. Thanks in advance!