imperva

REPORT

Lessons learned from analyzing 100 data breaches

Author: Ofir Shaty

Contents

INTRODUCTION

As part of Imperva's Research Labs' ongoing efforts to monitor and report on the current Database Threat Landscape we studied and analyzed over 100 of the biggest and most well known data breaches of the last decade.

To get the most value from the analysis of these breaches, Imperva researchers extracted each initial breach, the target of each breach, the system that was compromised by the breach, and what records were stolen in the breach.

Before we start with the lessons learned, it is important to understand that because of their sensitive nature, database security breaches often go unreported. Also, in many cases there is simply not enough data available to provide a comprehensive report.

The data we are presenting in this report has been gathered from all the available information sources; the web, breach reports, hackers' forums, analysis of stolen database dumps and information gathered from deploying honeypots. We present what specific types of data were stolen, the root causes of the breaches, year-over-year trends of the average volumes of records stolen, and which public cloud services are breached most frequently.

We also profile and characterize the types of attackers organizations face today and provide a flow diagram to help you understand the complicated ecosystem of data breaches themselves.

Our analysis of these data breaches reveals a bleak and troubling reality. Not only has there been a greater number and higher frequency of breaches over the last decade, but an increasing amount of the stolen data is being exposed and sold on the dark web. The stolen data sold on the dark web is being used by hackers in extortion attempts and as fuel to create "phishing" campaigns, which in turn leads to more data breaches.

The ecosystem of data breaches is varied and complicated. For example, many breaches are the result of poor security practices like publicly accessible service (Microsoft, Advanced Info Service) or weak authentication. Others are caused by unauthorized access of third party services (Cambridge Analytica) or supply chain attacks (SolarWinds, Accellion) . Some of them are turning a small scale breach to a colossal disaster like storing database passwords in clear text. Still others are caused by many other kinds of security practice failures, which we will explore throughout this report.

Inside threats and outside threats

To better understand the “Database Threat Landscape” it is helpful to characterize the entities from whom you need to protect your assets. Attackers are commonly split into two groups: “Inside Threats” & “Outside Threats”.

Inside Threats

Leave data exposed, either maliciously or by mistake/ oversight

Outside Threats

Attackers may choose attack methods depending on their motivation

Malicious attacks

Usually the motive is money, often accompanied by a dislike for the company. The malicious insider usually has access to assets, credentials and is most likely to be less suspicious than an outside threat.

Hit & Run

This “Opportunist” identifies an opportunity; whether it is a vulnerability, a publicly open database or something else. The actor decides to take what they can and leave. This kind of attacker will not try to search for other databases, or penetrate the organization’s network, they will not try to execute exotic exploits, etc. they will just take what they can and go sell it to the highest bidder.

The Curious

This attacker usually sets out with a purpose, but has some passion to look deeper to draw the line, they may look around a little bit, but not too much. They are still focused on their original purpose, malware deployment, data exfiltration, etc.

The Resident

The most dangerous type, as in the “Equifax” breach, the Resident will penetrate into the organization’s network and will stay for months, sometimes years. They will use keyloggers, sniffers and other methods to steal credentials and compromise databases, using “Slow & Low” and other methods to stay undetected.

Stolen data types

Hackers can only steal data that is available for them to take. In our research, we found that 74% of the stolen data is Personal Data. Over 15% of the breaches resulted in credentials being stolen and more than 10% resulted in the theft of credit card details.

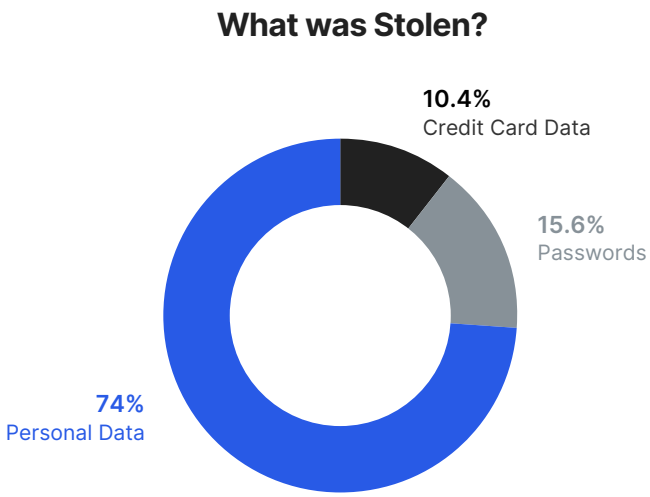


Figure 1: Distribution of Stolen Data.

Protect your Personal Data

Widespread theft of Personal Data is a strong indication that many organizations are simply not putting enough protection into place to secure it. Some of this is due to the fact Personal Data is regularly transferred between systems, people and suppliers to perform common business tasks. As regulations governing data privacy get tougher, it will be critical for organizations to discover, identify, and classify Personal Data across their data estate. Only when you know where the Personal Data is hosted and what applications and users are using it, will you be able to extend the security controls that protect it.

The data suggest that credit card details are the most vigorously protected data set of the three, but also more intensely sought after as it is in high demand on the dark web.

Root causes of breaches

Imperva research shows that almost 50% of the data breaches begin in the Web Application, either through a SQL injection (SQLi) or some other vulnerability like Remote Code Execution (RCE) or a simple information disclosure. The Web Application is the front door to an organization's sensitive data and a reliable Web Application Firewall can mitigate these vulnerabilities and limit their exploitation.

Another big trend that stands out is the 15% of data breaches whose root cause is listed as “Publicly Accessible”. A publicly open service increases the risk for a breach to happen, but in most cases this is not a failure of security practices, it is rather a complete absence of a security posture. For example, when there is a failure to secure non-production environments where company data may be used to perform testing. We’ll offer a deeper analysis of this part in the next section.

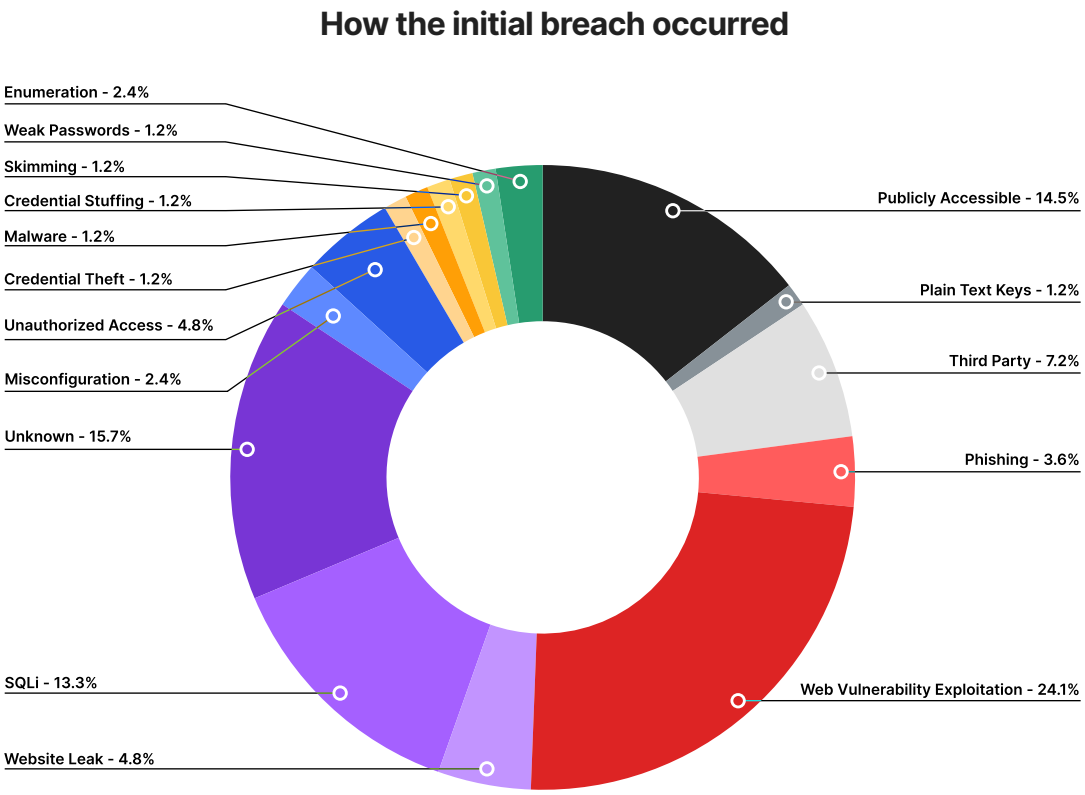


Figure 2: Distribution of Initial Breach

The human error factor

Some data breaches are a result of human errors. Privileged users’ mistakes or misconfigurations, either unknowingly or through negligence, lead to data being left accessible or leaked.

The threat may come from within the network space, but can also be in the physical space; the theft of drives, laptops, or printing the data instead of transferring it via a network.

Public cloud services

Data breaches in open public cloud services are trending dramatically higher as more and more companies are migrating their operations and workloads to the cloud.

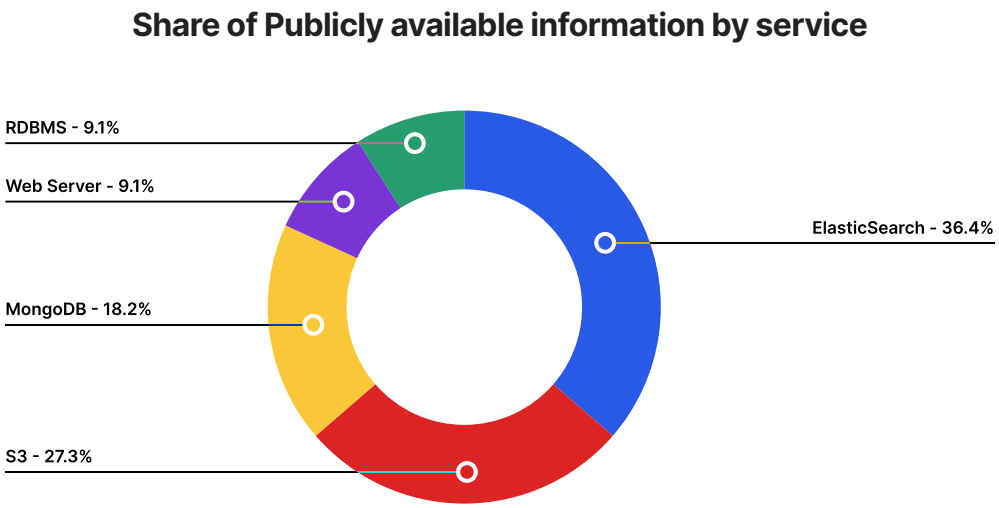


Figure 3: Distribution distribution of publicly open services

In the chart above, we see ElasticSearch as a trending publicly open service with 36.4%. By nature, ElasticSearch contains big amounts of data to perform search and analysis. Since it's a search tool and not commonly defined as a "data store" many organizations may unknowingly leave the system open to public access. Public cloud services breaches may also be the results of cloud configuration issues, vendor defaults, etc.

Second behind ElasticSearch is the Amazon S3 bucket with 27.3%, S3 buckets are commonly used as a data store for data lakes.

This is not surprising at all, the publicly open services just sitting out there are easy prey. There are multiple ways to target public services, from Shodan search to open source apps like LeakLocker that provide accessibility for the hunt.

Average number of records stolen per year

Our research shows an increase in the volume of data stolen every year. In 2020 we started to see more and more breaches that exfiltrate records in billions.

According to the analysis of thousands of data breach details published on dbdigest, we made calculations on the raw data and found some interesting information about data breaches.

Average number of records stolen per year

Year	Data breaches	Number of records compromised	Average record compromised per data breach	Increase in breaches	Increase in records stolen by year	Increase on average
2017	488	826,526,181	1,693,701			
2018	557	2,340,329,140	4,201,668	14%	183%	148%
2019	956	12,304,182,843	12,870,484	72%	426%	206%
2020	1,120	20,212,424,547	18,046,807	17%	64%	40%
2021 (Jan)	**82**	**878,168,975**	**10,709,377**			
Average		8,920,864,678*	9,203,165*	34%	224%	131%

* through 2020

Figure 4: data breaches, records compromised and average record compromised per data breach per year.

- In January 2021 alone we saw more than 870 million records compromised. This is more than the total compromised records for the entire year in 2017.
- We see a constant increase in the number of data breaches of more than 30% each year.
- Another interesting piece of data is the number of records compromised in each year, again constantly increasing, by an average of 224%.
- Based on this trend we can estimate that year-over-year we will see around 3 times more records stolen annually.

When we put this data from the chart above into the below graphs, we can see a strong correlation between the three data points. The data breaches count is represented by the red line, the count of the total number of records compromised is represented by the yellow line and the average number of records compromised is represented by the blue line.

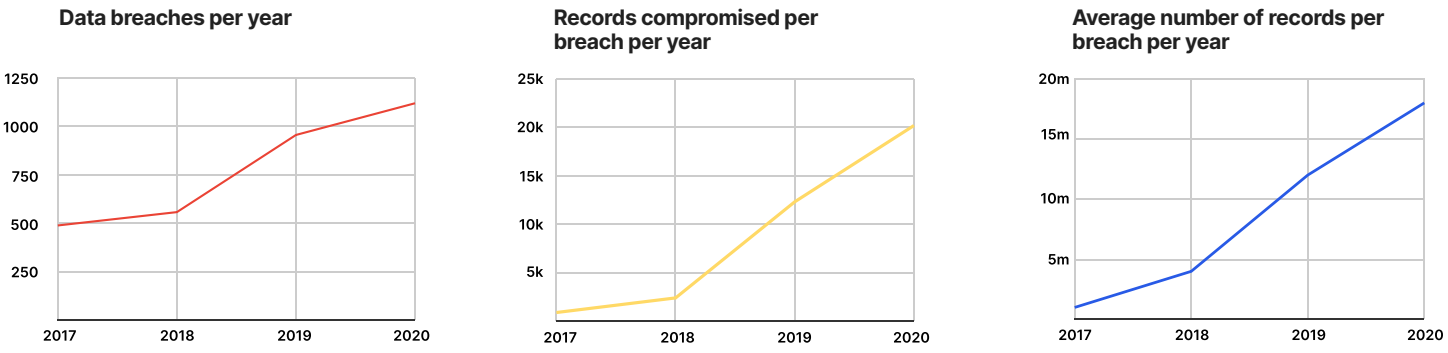


Figure 5: Line graphs showing the number of data breaches per year, the count of the total number of records compromised per year and the average number of records compromised per breach per year from 2017-2020.

We predict that for 2021, there will be around 1,500 data breach incidents with a total of 40 billion compromised records, and an average of 26 million compromised records per breach.

The constant increase in data breaches is a result of multiple factors. We are living in a digitalization era in which more services are consumed on a daily basis with the majority of them online. More businesses are migrating to the cloud which makes them more vulnerable if not done carefully. The increase in the amount of stolen data is the result of similar factors. The amount of data that is out there is enormous, and it is increasing every year.

Information security adoption is slower than the adoption of digital services that make profit from the addiction to and consumption of the same online services. The increasing number of breaches every year is a result of this gap.

2020 was a year with a big impact on digitalization, with many sectors making a very quick shift into digitalization to make themselves available through the COVID pandemic. Such a fast, dramatic change is likely to have security implications.

Use the Data Breach Matrix

We created flow diagrams to help you better understand the complicated ecosystem of data breaches.

Most data breaches are actually a series of failure points rather than one specific activity that results in a breach. Looking at attacks this way helps us understand the timeline of events, from the initial breach to the actual exfiltration of records.

Databases may be used as a target point for “Data Exfiltration”, or as a penetration point into the organization network to get a foothold, deploy malware, etc. Those are two different scenarios, but there is an overlap and common areas in between.

Consider the scenario of Information Disclosure of credentials within a vulnerable web application together with a publicly open database server.

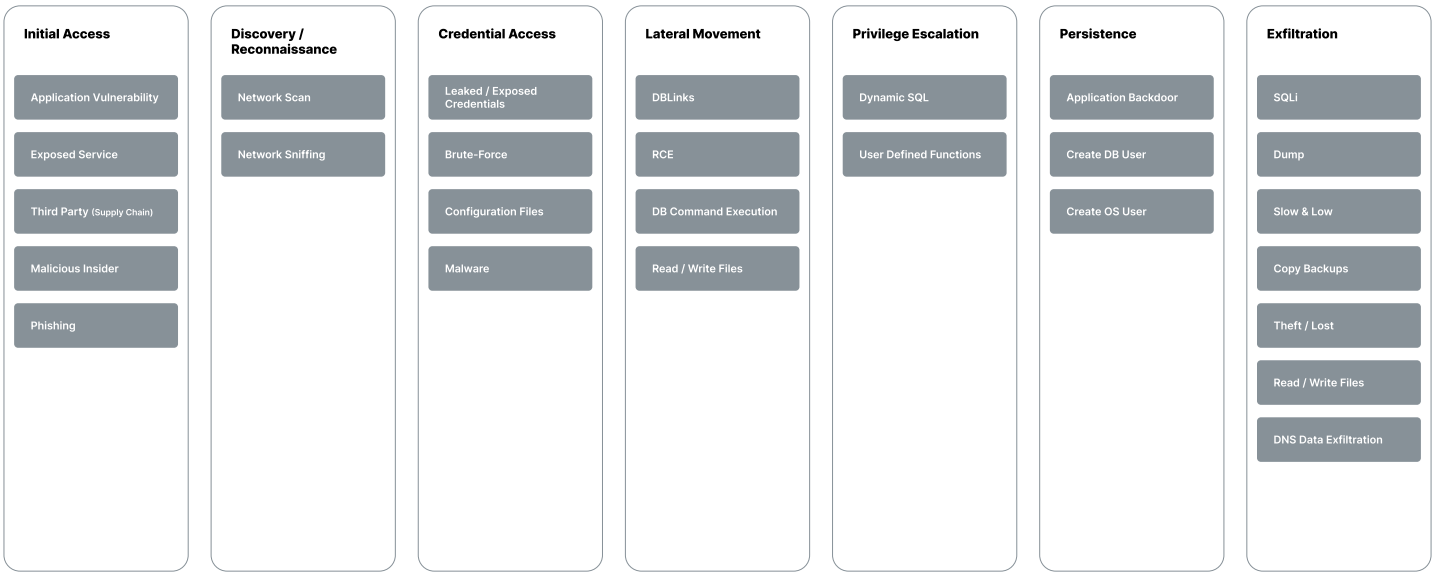
A database with stored credentials that are not encrypted or poorly encrypted can be a fertile jumping point to attack other databases in the organization and can lead to enormous breaches where dozens more databases and the sensitive data within them can be compromised.

The stories are spread from public Github repositories to private S3 buckets, from RCE to Ransom or from simple and old Web Vulnerability to a massive crypto mining botnet.

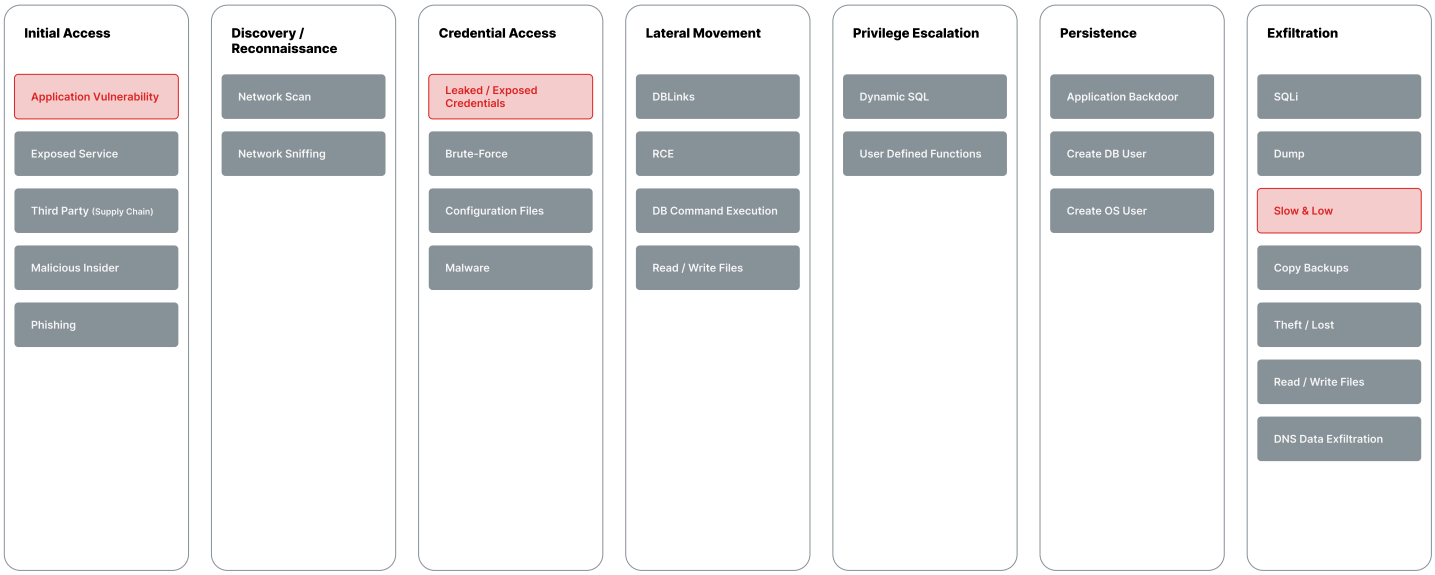
Figure 6 below is the data breach matrix we created. It describes the flow of a breach throughout the kill chain steps, from the initial access until the exfiltration.

Every step in the kill chain can be achieved by different techniques represented as blocks inside each kill chain column.

The data breach matrix



The following illustrates the flow of the Equifax data breaches within the matrix:



Summary

Traditionally, we've believed the assets within the network perimeter were protected. But as services are digitally transformed, the boundaries of the perimeter have blurred. **The security of an organization is only as strong as the weakest link in the security chain.** Many times, the "walls" that protect databases have cracks that allow attackers to put their hands on sensitive data. In many cases, better architecture and cross organization security practices would do the trick, but those practices are not easy to implement and control. We suggest that organizations implement security for the databases they manage, not just the applications and networks that surround them.

In most of the analyzed breaches the lack of in-depth security stands out as a main reason. Organizations can reduce the attack surface by securing their database environments. Since a significant number of the attacks target web applications, separating the database server from the application server can make a big difference. Together with dilution of excessive privileges from key users and strong authentication mechanisms, these practices in combination can help you avoid a data breach.

Protecting your organization's data is a never ending process, you must always work toward optimizing your security architecture, policies and practices, both for your assets and employees.

Continuously performing "Discovery and Assessments" (**DAS**) to locate the sensitive information and find security holes is a great way to stay on top of your organization's security posture and eliminate bad practices inside the database environment.

Together with implementing security products like "Web Application Firewall" (**WAF**), "Database Security" and "Database Risk Analytics" (**DRA**) it is possible to protect against most of the scenarios presented here.