DATAGOV

KIERAN KEENE

PREFACE

This book is intended for anyone that wants to venture into the world of data governance. This book provides some of the best practices I have discovered throughout my career and is not intended and should not be used as legal advice for your GDPR compliance. For this, it’s always best to seek legal advice.

INTRODUCTION

Information governance is an incredibly important topic in the ever expanding world of Big Data. If we are honest with one another, it's not the stuff dreams are made of and it's unlikely to be anyone's dream job. It is however, critical.

Information governance is all about managing data to keep it safely tucked away in our systems and away from unauthorised individuals. To do that, we need to put systems, processes and policies in place to keep our customer data secure and to protect the reputation of the business.

We have seen a lot of data breaches in recent times, even big companies with departments dedicated to the task of protecting customer data are falling victim to large scale data breaches - damaging customer trust in the brand and incurring very large fines. Talk Talk; British Airways; and Yahoo are some examples of large breaches in recent times.

We are never going to stop data breaches, but we can work to reduce the frequency and limit their impact by implementing appropriate governance processes and policies and by driving a data culture within the business - holding everyone accountable for the safety of your customer data.

Making employees accountable may sound like a bizarre approach to data governance, because it's all system-driven, right? We just need an additional layer of security, some encryption and we'll be ok, correct?

No. Kaspersky reported in 2019 that 90% of data breaches in the public cloud were a result of social engineering, which is all about using deception to trick individuals within a company into divulging information that they shouldn't. An example may be a phishing email, which prompts the user for a password on a fake website, giving the hackers direct access to the data.

So indeed, the human element around the systems are just as important if not more important than the technical solutions that we use to secure our data.

Through this book, we will talk about ways to prevent data breaches happening; protecting your customers and brand. These methods will fall under four main headings: laws & regulations; people; processes & policies and technical security.

From all this information, you should be able to start to put together a data governance strategy for your business and work towards a well governed environment.

LAWS & REGULATION

Data protection laws are put in place to protect the data subject, which is the person that the data is about. For example, if we had a record that had my name, phone number and address, I would be the data subject.

It's all about protecting my PII (Personally Identifiable Information). That could include my full name; address; email address; national insurance number; passport number; credit card number; date of birth; phone number and plenty more information.

We also have what we call 'related' information. This is where, the single piece of data would not be enough to identify me as a person. For example, knowing that I am between 25 and 30 years old, is not enough to identify that the data record relates to me, but if you were to link that to other information, such as gender, job title or city, you would know that you'd be looking for a male Big Data Architect in Woking between 25 and 30 years of age. It then becomes possible to identify me as an individual.

If we think about it in the telecoms arena. Knowing someone's device is a Blackberry Curve 9100 is not personally identifiable information. But, knowing that in conjunction with their age band and city definitely is. There aren't that many of those devices left, so there may only be one 25-30 year old in Woking using one of those devices & therefore it may be possible to distinguish them from the rest of the customers.

Some examples of relatable information would be:

- First or last name. This would only be if those names were common, like Ben or David. If you have a very uncommon name, it would be PII.
- Country, city, postcode
- Current employer and job title
- Gender
- Race

Finally, we have data which is definitely not PII, which can include IP addresses and cookies.

It’s important that we categorise PII data correctly. If it is personally identifiable, we need to treat it differently to non PII data. We don’t really need to encrypt Non PII data as a customer would not be identifiable, should there be a breach.

What is the main regulation I need to worry about?

This section will provide an overview of GDPR (General Data Protection Regulation), which will give us the principles / guidelines that we need to follow to make sure our customer data is safe.

The GDPR regulations were introduced in May 2018 to replace the Data Protection Act, which was written in 1995. As you can imagine, with rapid advances in the way we are using, sharing and storing data, a data protection policy written 23 years ago isn’t going to cut it.

The purpose of the policy was to give European citizens more control over their data and to prevent companies from misusing that data for things that the data subject (the person the data is about) doesn’t agree with.

At a high level, those guidelines are:

- Process data lawfully & fairly
- Collect data for a specific purpose
- Only keep what you really need
- Store the data securely
- Hold people accountable

Let’s dive into each, starting with **processing data lawfully and fairly**. This all starts with the data subject giving consent for you storing and using their data. Without this consent, you can’t store it or process it for any reason.

You have to be absolutely transparent with the user and have to tell them exactly what your intentions are with their data. You need to tell them things like: exactly how are you going to use that data?; are you going to share it with third parties? And how long are you going to store it for?

You are also committed to **making sure that data is accurate and up to date**. If you ran a marketing campaign last year and you know some of the email addresses on your list were incorrect, as people complained. You must rectify or delete those email addresses as you cannot store data that you know to be incorrect.

You must also give the data subjects the means to request that their data is removed or updated to be correct. If such a request is received, you must do it within one month.

You must also only collect personal information for **a specific purpose** and only retain it for as long as it’s supporting that purpose. For example, if you were Facebook, the specific purpose of holding your information would be to provide the service that you signed up for. Without your email address & password, it would not be possible to login. Without retaining your phone number, the forgotten password process wouldn’t work and without storing your full name, you friends would be unable to find you.

If however, I was a company like a local grocery, there would be no need for me to collect your phone number and email address, before selling you a broccoli. Therefore, collecting this data is not supporting the specific purpose of providing you with the service you requested.

Then it comes down to **only storing the data that you really need**. First, you need to ask yourself about retention. Do you really need to retain the details of someone that closed their Facebook account 5 years ago? Probably not.

We then need to question whether we actually need all the information we hold. For example, does holding my gender, race and income support you in providing the product or service to me? If you're a bank, probably - they use this information in machine learning algorithms to predict the likelihood of defaulting on loan payments. If you're a personal fitness trainer, you probably don't need it.

It's important to delete the data that you don't absolutely require to reduce the risk exposure that you have and reduce the impact of a potential data breach.

Now we get to the technical element of GDPR - **storing data securely**. Let me give you an example of how not to do this. I used to do some work for an online company and they stored all of their customer and order details in a spreadsheet. That would be fine, but that spreadsheet was on a server, which had no authentication and was indexed by Google. So, all customer details were exposed to anyone that took the time to try and find them.

Clearly, that is an extreme example but it's a true story and these things do happen. It is our responsibility to make sure that we put appropriate security provisions in place to stop breaches happening. These could be simple mechanisms, like enabling multi-factor authentication and storing our data on Google Sheets (rather than the dodgy web server I described above), or we may choose to implement more complex solutions.

The key thing here is that it must be a reasonable provision. Password protecting those spreadsheets on the server would not be reasonable; it wouldn't be much more secure than having them passwordless. Equally, investing in a £500,000 system to manage those spreadsheets would be unreasonable from the company perspective - a huge cost. There is a balance.

Finally, we need to **hold people accountable**. In my mind, this is a combination of the company as a whole, where we must document exactly how we're using and managing customer data, but also we must drive a data culture in the business, upskill our staff and hold them directly accountable for the safety of customer data. Ultimately, the more individuals in the business that can demonstrate your compliance to the GDPR regulations, the better position you will be in.

Money, money, money

Beyond just doing the right thing for your customer and making sure their data is safe, we also have an additional motivator in the form of huge financial fines. These can be either:

- Up to €10M euros or 2% of annual turnover (whichever is greater)
  OR
- Up to €20M euros or 4% of annual turnover (whichever is greater)

Which fine you receive depends on what you've been perceived to have done wrong. Below are a few examples of things you can do if you want the biggest fine possible:

- Are not processing data lawfully
- Don't have consent from the data subject
- Transfer data to third parties without the data subject agreeing

The difficulty with GDPR is, you cannot possibly estimate the value of your fine. Firstly, there isn't a huge amount of precedent yet – we haven't seen that many organizations being fined… yet. The second reason is that there is not absolute criteria in place. The behaviour and intent of the organization will be taken into account when calculating the value of the fine.

What does that tell us? It tells us that the governing bodies are well aware the data breaches happen and companies can't always stop them. So when they investigate, they take into account our intent to comply to regulation.

- Did we drive a data culture within the team?
- Did you have appropriate mechanisms in place to identify and report the data breach as soon as possible?
- Did you have policies and procedures in place to manage data appropriately?

It’s fair to say that you’re still going to get a fine. But, if you’ve done everything in your power to prevent a breach it should be lower than if you’ve had blatant disregard for the safety of your customer data.

If there is a data breach in your organization, it is mandatory that you send a breach notification to all impacted parties within 72 hours of having become aware of the data breach. So again, having the processes in place to identify breaches early, will lead to a better standing with your customers and the regulatory body.

Data subjects have more rights too! They have the right to access & the right to be forgotten. The right to access is the right to see what data the company holds about them (which must be done free of charge, in electronic format) and how that data is being used and processed.

The right to be forgotten is the right to request their data to be erased and processing halted. The user can only request data to be erased under certain circumstances though, including: if the data is no longer relevant.

It’s important that you don’t fall into a trap around these regulations. There is a misconception that US businesses don’t need to adhere to GDPR as it’s a European regulation. This is however untrue. It’s very likely that you hold at least some data on your servers that belongs to an EU citizen and as such, you must comply to the GDPR regulation, or you could find yourself in a world of pain!

If you don’t have no intention of doing business with an EU citizen, you can use rules to drop requests from EU countries on your website, so you don’t risk the users creating an account or filling out a form. Of course, doing that could damage your reputation for US customers that have travelled abroad and are unable to access your website, but it is certainly a strategy which could be adopted.

So let’s say then that you are going to be dealing with EU citizens, how do you get their consent to store data? Well, it had to be an affirmative action. That means, the user has to acknowledge that you will be storing their data and agree to it – you cannot pre-tick the checkbox or implement other workarounds which mean the user passively agrees.

You must also keep a log of how and when the person gave their consent and you must respect requests to withdraw consent immediately.

So, who’s been fined?

At the time of writing, the GDPR regulation hasn’t been in full force for very long, but we already have some big names with some very big fines:

- British Airways:	£183 million
- Marriott Hotels:	£99 million
- Google:	£50 million

The list above shows, GDPR is no joke. These fines are probably relatively insignificant to the likes of Google, but small companies are getting hit with several thousand pound fines which could kill their business.

How can I avoid those fines?

To avoid the fines, you need to adhere to the regulation 100%. In addition to those regulations, it’s also important to have some best practices within the business.

First off, you should only give people access to the data that they absolutely need to carry out their job. For example, Sue in retentions absolutely requires access to customer records so that she can service the customer when they call. Robert in Finance doesn’t need to look at customer details, so he should not have access to do so.

It sounds simple, but very often I see access requests coming through saying ‘I need the same access as Bob’. The thing is, you don’t do the same role as Bob, you just know he has access to the system and you don’t really know what you will need, so you want everything you can get. It’s not a valid request and that sort of thing should be challenged.

Additionally, just because Bob needed access three months ago, doesn’t mean that he still does. His access requirements may have changed and we need to carry out an audit to validate if changes need to be made.

We will talk about some more of the policies, processes, controls and technical security that you need to implement to keep your customer details secure in subsequent chapters / sections. This isn’t supposed to be an exhaustive guide or legal advice but hopefully it will support your GDPR compliance.

PEOPLE, PEOPLE, PEOPLE

In the last chapter we talked about why we protect our customers data. It’s all about adhering to regulations and keeping out customers safe. In this chapter, we’re going to start talking about how we go about effectively governing our data, specifically, we are going to talk about the people aspect of data governance.

You may remember from chapter one that people are the root cause of 90% of data breaches so having training, committees and specific data-related roles within the business is vital.

Let’s start by talking about roles, starting with the data owner. Their responsibility is to make sure that the data they own is governed across the business. This can take many forms, but in short, they need to:

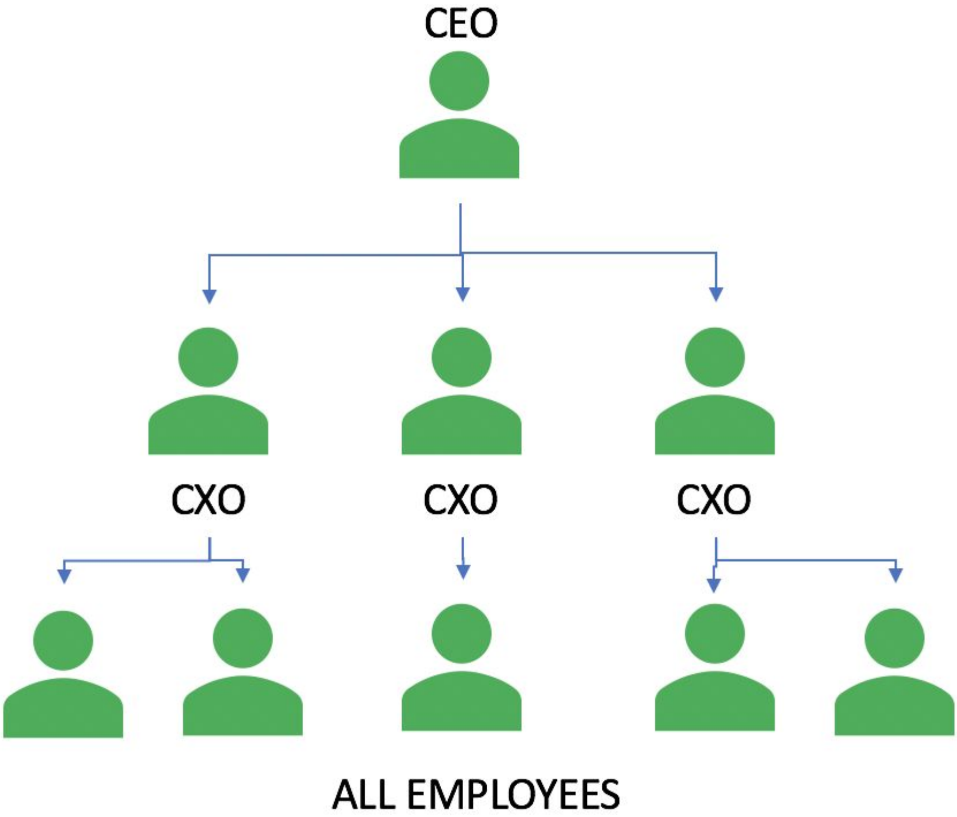
- Retain a full data dictionary, defining the data they own.
- Ensure data accuracy through validation efforts.
- Document and manage the lineage of the data (what processing has been done to the data to get it to its current point).

Ultimately, the data owner is fully accountable for their dataset and the individual should be senior enough to carry significant weight when issues are raised. A data owner would be supported by one or many data stewards.

Data stewards are the folks that have the accountability for all the day to day management of data. They’re the people that are absolute subject matter experts and can advise the business around the data that is held.

Finally, we have committee members. The committee are an advisory board for the business. They are responsible for defining and clearly articulating the data governance strategy and guidelines. They are responsible for championing data management across the business and make sure that regulatory demands are met.

If we think about the structure of a business, all the people in green are responsible for data governance.



Yes, that's right! Every single member of staff, regardless of their seniority should be accountable for our data governance. Why? Because, we all have the ability to influence this in some way. If you work in building maintenance, it's true, you may not have access to all the IT systems, but you have access to the building. You still need to make sure that nobody is tailgating you through the door and that you don't allow unauthorised individuals access to the filing cabinets.

A big part of this accountability is training. How can you possibly expect your employees to support your data governance efforts effectively without investing the time and money to train them properly.

Data security policy

People are the weak link when it comes to data security, so we need to do a number of things to help them to help the company and protect your data. One such deliverable is a data security policy. This is the bible as far as data security is concerned and it'll help your workforce to do right by your customers and their data.

Within our data security policy, we need to give you're an employees an overview of why they should care. It should talk about the possible impact to the customer, the regulatory requirements and it should talk about what is and isn't an acceptable way to process customer data.

We should then cover off all sorts of things which will help safeguard customer data. The first, is to create a password policy for your business. If John has a password of John1, he is asking to be hacked. So, we should force our employees to have passwords with: upper and lower case letters, special characters, numbers etc… something like this: '_GSE+QW5a+e3NcDy'.

Passwords are a huge problem in businesses, did you know the most common passwords are 'letmein', 'password' and '123456'? When we have phrases, sequential letters or something that relates directly to the user as the password, it's like giving the hacker the golden key – it won't take them long to break it. Whereas, our super complex password above, will give them a lot more to think about.

The next part of our policy needs to surround internet usage. We need to make sure that our employees know that misusing the internet can lead to data breaches. If for example, your employee were to download a virus to your corporate network, you'd be in for a bumpy ride! Defining what they can and can't do and keeping a log of those activities is important for your information security.

We need to also define an email usage policy and train our users on Phishing. As with the web usage, it's perfectly possible to accidentally download a virus from an email. If that happens, we need a clear process to declare the problem to IT and in turn IT need to know exactly what to do to minimize the impact.

With regards to Phishing. This is where the hacker is fishing for information. They might send an email claiming to be from a supplier or other company and try to draw important information from you. This can be conducted over phone, SMS or email, so you need to train your employees and document what to do when they suspect phishing.

The next key part to the process surrounds mobile devices. It is the employees responsibility to make sure that their device meets the security criteria set out by the company. For example, if they are using their devices for work related matters, there should be password protection on the device; relevant security software; the software versions should be kept updated to keep up with bug fixes and plenty more of other criteria they need to ensure they meet.

Finally, for any breach of security, you need to define a process to report them so that the impact can be assessed, and the company can take steps to mitigate the damage.

Training and engagement

You probably agree by now that the single biggest threat to a company's data security is it's own workforce so we need to train them and ensure that they understand the implications of a data breach and what they can do to mitigate such a breach.

The problem is, this kind of training is boring. Really boring. So, we need to find ways to actually engage our employees, rather than simply ticking a box because they said that they read the policy.

I’ve seen this engagement done well in the past. I’ve also seen it done very badly. A few ways I have seen to really pique interest are:

Running Data Days: this is all about showing the value that can be derived from data when it’s handled correctly. You can get your workforce excited about the potential to automate their currently very manual reports; to gain additional insight about their customers or to run machine learning models across the data to make predictions as to whether a customer is likely to respond well to an upselling attempt by the sales team.

By showing the possibilities of the data, it does two things. Firstly, it shows the team the value of the data to the business but it also shows the level of sensitive data that you store about your customers – this is something that many people in your team may not be aware of as it doesn’t directly impact their day to day job.

As part of the day, you can run engaging data governance workshops. You could show some ridiculous phishing emails (like the one from a foreign millionaire who just wants to transfer some money out of the country and needs your help to do so. Don’t worry, you’ll get a 10% cut), coupled with some more serious ones – make a game of it, perhaps you could call the session ‘to phish or not to phish’ to pique some interest on the agenda items too.

There is no escaping some sort of formal training and assessment too. It’s the only way to truly validate that your team were paying attention and that they understand their responsibilities towards your customer data. It also provides you with an audit trail. Remember we said earlier that if you have tried to comply with regulation, it may aid you in a lower fine than if you've totally disregarded it? Well, showing that you have trained your employees and were fostering a data culture is definitely a positive thing to show the regulators.

The training can be delivered as an online course with quizzes or as a face to face session. My preference is a face to face as you can get people up out of their seats, interacting and can make the session a lot more engaging. Conversely, should you provide an online training session, you'll probably find that your team are not listening and are rather surfing the web or responding to emails. As this is a really impactful and important training session, you should make as much effort as possible to deliver it as a face to face session.

Next is certification. The promise of an industry certification is very good for focusing the mind and getting people on board with the programme. It's not feasible to put everyone through formal training, but you can have a champion in each business area that can advise their colleagues and help foster the data culture you need.

It's important to keep the momentum. You can do this by running weekly cybersecurity drop-in sessions or running working lunch sessions where you talk about interesting things in the world of cyber security. The key thing is to make sure that you make a conscious effort to make the content engaging. Death by PowerPoint does not drive a data culture, whereas running byte-size cyber security training sessions, which allow the employees to work towards a professional accreditation would. There has to be something in it for them too!

POLICIES

One of the key parts of any information governance strategy is policy creation (and of course enforcement). It comes back to this notion of showing you tried to protect the customer data and failed is better than not trying at all.

With data, it is important to have the below policies in place:

- Data retention, archiving and disposal policy
- Data classification policy
- Information sharing policy
- Data quality policy
- Subject Access Request (SAR) policy

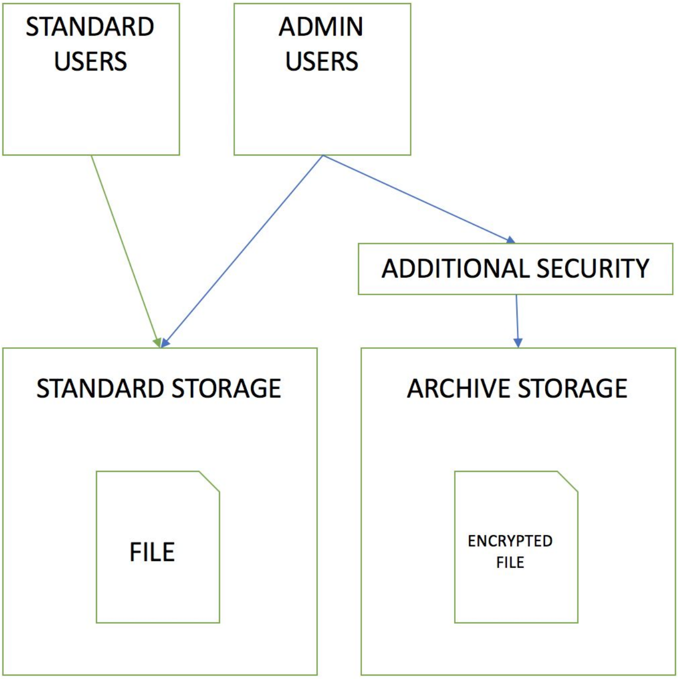
Let’s start with the data retention, archiving and disposal policy. This is all about how long you need the data for. The below is an example of how that might work.

Database Table	Archive After	Delete After
table1	30 days	90 days
table2	360 days	7 years

The idea here is that you need ‘immediate’ access to the data within table 1 for 30 days. After that time, it doesn’t really need to be accessed very often (if ever) and can therefore be archived, but must be retained for audit purposes.

The benefit of this is, we can have even fewer individuals with access to the archived files and we can add additional encryption to the files. Even if it takes some time to access the raw information, it’s not a problem - as long as we have instant access to the newest data.

So our security increases on historical files. In the below, we can see that once a file moves to archive storage, standard users no longer have access to it - only admins can access the files. The file has also been encrypted and we have an additional layer of security to prevent unauthorised access to the file.



These things can be systematically done too. For example, with AWS (Amazon Web Services), which is a popular cloud provider, you can automatically send files to an archive after X days and you can automatically delete those files after a further X days. This takes any human error out of the equation and ensures that your data governance policy is actioned.

One final observation is that each file can have a different ruleset. For example your accounts may need to be retained for 10 years, so your archiving and deletion policy would reflect that. Whereas, your customer order information may only be required for 1 year and hence, you can drop those files sooner. Each file or datasource should be treated independently and have their own archiving rules.

Next, we have the data classification policy. This is a really important piece of work that organizations must carry out. It's all about classifying how important data is and hence, you can decide how to protect it. Many organizations will have a scale, like the below:

- Public: this kind of data is available to anyone, for example, the companies quarterly accounts are available on the company website for anyone to download and consume. Hence, we should consider these files low risk.

- Internal: this data is slightly more sensitive. It could be that these are the yet-to-be-released quarterly accounts. So, these documents should not be communicated outside of the company. The security policy should reflect that.

- Confidential data is stuff to do with our customers, like their name and address. This data is super sensitive and we need to make sure we protect it at all costs. This data should only be accessible by people that absolutely require it and the security around the data should be extremely tight.

- In strictest confidence data would be the next level of sensitive information. It could include biometric information about our customers (e.g. fingerprints) or their payment details. This data should be accessible to even fewer individuals in the company and should have the tightest controls around it.

This policy may seem like an administrative headache, but it's absolutely necessary. We can use this policy to drive our investment into our technologies to protect customer data; it also tells your employees what they can and can't do with that data.

That is a very important point. Just because we may think it's obvious what classification a piece of data should have, it is a subjective thing. So, your workforce might consider a file to be confidential whereas, you consider it to be the next step - in strictest confidence. We can remove that subjectivity and confusion by tagging all data with their classification.

The information sharing policy is exactly what you'd think. It tells our workforce what information they can share and with who. This includes sharing information internally between teams but also how they communicate with third parties.

If you have a third party service provider that supports your business, can you share customer information with them? Well, read your information sharing policy to find out!

The data quality policy is important too. We need to make sure we keep accurate records about our customers. If you think about it from a healthcare perspective, if we see that someones blood pressure was 15,000, we know it's inaccurate. These inaccuracies do however, occur - all roads lead back to our good friend, human error.

We can therefore implement some systematic checks and prevent people creating records with incorrect data - we call these domain restrictions. For example, in a hospital, you might set the age field to be anywhere from 0 to 120. In the event that someone mistypes 40 as 400, it will not allow them to create the record.

We can also implement regular data audits to check the distribution of the data. Where we see extreme outliers, we can have those flagged automatically and we can follow the process defined within this policy to rectify the data inaccuracies.

We can also look at the consistency of our data. What does that mean? Well, it’s very likely that the same piece of data resides in multiple places. It’s terrible but it’s true, it almost certainly does. So, if I have a system that stores the patients blood pressure, I can validate the data quality by making sure that it matches the other systems that store the same data.

We can also look at how many incomplete records we have. We need to ensure both the accuracy and completeness of our data. We therefore need gaps to be flagged and we actively need to work to populate the missing data. If you were a nurse, you could do that the next time you see the patient.

Finally, if we see that we have bounced emails, where the email address is not valid, we should reflect that in our system. We should then contact our customers via other means to fix the data at source and make sure it’s accurate.

The subject access request (SAR) policy is all about how we provide our customers with access to their own data. This is a complex policy as we need to make it easy to find out where all the customer data is stored. In large corporate businesses, this could span tens of systems and be quite difficult to collate.

This policy needs to outline how to obtain all the information but also, what criteria the customer must meet in order to be provided the information.

TYPES OF ATTACK

Before we get into the technical solutions that can help protect against security breaches, let's talk about the most common types of hacking. Once we know what they are, we can protect against them.

Malware is probably the type of cyber security breach we hear about the most. We have some huge examples of this with WANNA CRY among others recently. The malware can get onto the servers through social engineering methods (discussed later), where a phishing email or convincing advert leads to the user clicking on a link & downloading the dreaded malware.

One thing to really think about here is, if the hackers manage to obtain the password to one of your systems, it's very likely that your users have re-used their password on other systems too. So it can have a trickle down effect and leave you very exposed.

Another example of getting malware onto your system can be through the use of a fake wireless access point. The hacker could setup a wireless router somewhere near a Mcdonalds restaurant and call it McDonalds1. This would trick users into joining the network, thinking that it's legitimate and managed by Mcdonalds themselves. Once connected, this can lead to malware being downloaded and could lead to sensitive information to be stolen. This absolutely therefore requires you to include a section in your policies about joining networks that you don't know.

The next thing we're going to talk about is SQL injection. This is a type of attack that can be used with web applications using SQL as the backend database (or MySQL, PostgreSQL etc..). Essentially, if the website is not properly secured, hackers can insert (or inject) code into the website, enabling them to extract the username and password.

This in itself will be impossible for your end users to spot. So, it’s important that they are only logging into secure and properly managed websites, rather than relatively unknown websites. Again, it’s likely they re-use their password, so getting hacked on www.i-love-to-buy-brocolli.com may not seem like an issue, but it is if they’ve used the same password for your CRM system.

One of the most painful parts of dealing with hacking is realising that most of it is the fault of people. Social engineering is what we call a type of attack where users are manipulated into believing a message is from a trusted source. From here, you could download an email attachment, get malware on your PC and suddenly be compromised.

Some types of social engineering include:

- Phishing. This is the most common type of social engineering attack and I actually received one recently. The hacker had recreated an exact replica of the kind of email I might receive from my bank, they had also setup a fake login page, so if I had taken the bait, I would have just given them my username and password. Of course, I don’t believe anything I am emailed and checked the email address it came from, clearly it was not from my bank - you would be surprised how many people do fall foul of this,.

  There are different types of phishing too. It can be over email, SMS and the phone but it can also take the form of online ads, where they entice the user to click, thinking it’s a brand they know that is having a 50% off sale. Once you’ve clicked, they prompt you to login on a replica login page.

  Spear phishing is another type of phishing where the emails are a little bit less random and generic. They’re targeted at particular individuals and they’re made to be extremely unique. Imagine they did a lot of research about you online and could tell you all sorts of things they shouldn’t know about you. You’d be more likely to fall into their trap.

- There is also Vishing. This is where the hacker creates an IVR (Interactive Voice Response) system that is identical to that of a company you know. They then trick people into entering their confidential information into the IVR when requested.

- Baiting plays on the curiosity of people. Attackers will leave a USB drive on the desk and wait for someone to come along and stick it in their computer. The USB drive has malware on it and infects the users computer.

  More modern versions of baiting will be around downloads online. For example, you may see if a user will take the bait of a free software download.

- Tailgating is a type of physical security risk. When you enter a building that requires a pass to get in, unauthorised individuals might tailgate you and get through the barriers before they close behind you.

We may talk about hacking a lot, but it actually doesn't account for that big of a chunk of attacks. Hacking is all about finding and exploiting vulnerabilities in software. This attack allows the attacker to perform actions on the system that they should not be authorised to do.

We see apps being updated all the time to fix bugs. It's really important that we keep our apps up to date, as those bugs may present vulnerabilities in the application, which hackers can exploit.

In recent times, hackers have taken control of systems in major organizations and taken actions to extort the companies out of money to restore the system back to it's working form. In the Netflix case, they then released the unreleased episodes of Orange Is The New Black, because Netflix didn't pay them. Imagine if those episodes were your customer data!

Another issue is when the user credentials to various systems get cracked by the hackers. This is a particularly big issue, as users tend to use the same password for multiple things.

As a result, more and more people are turning to password management software, so they can have an increased number of more complex passwords. The issue is, the password managers are software which will inevitably have vulnerabilities, which can be exploited and all passwords can be exposed at once.

The final type of attack we will discuss is DDoS which stands for Dynamic Denial of Service. This is where the attacker floods the server with requests which overwhelms the systems and prevents legitimate requests from being fulfilled. There are lots of workaround for this problem, which we will discuss in the next section.

So the conclusion then. The below table shows each of the attacks we described above and looks to the most suitable solutions.

Attack Type	Solution
Phishing	- User training - Security policy - SAR Policy - Information Sharing Policy - Technical solution (e.g. spam filters)
Vishing	- User training - Security policy - SAR Policy - Information Sharing Policy - Technical solution (e.g. blacklisting numbers)
Baiting	- User training - Security policy

Tailgating	- User training - Security policy
SQL Injection	- User training - Security policy
Hacking	- User training - Security policy - Technical solution
DDos	- Technical Solution

The above table really shows the problem here. Of course, we need the appropriate technical solutions to manage our data, but people really are the problem in every business. We’ve already discussed that they are the root cause of most data breaches and hence, we need to make sure our policies and procedures are absolutely bulletproof in addition to good technical solutions.

For Phishing and Vishing, we can put technical solutions in place but it’s likely that some messages or calls will slip through the cracks. The employees must therefore be trained to identify those mails that we would consider to be phishing; they must know how to keep themselves safe (security policy), they then also need to know what they can share and with who (SAR and Information Sharing Policy), incase they are fooled by the email and share the wrong information.

Baiting really just requires the workforce to agree to the security policy and undergo the associated training to make sure that they’re not tempted to put that USB stick into their machine.

Tailgating needs the workforce to be trained to identify tailgating and they need to act to stop it before any harm is caused.

When we think about hacking, we need to train the users in-line with the security policy to make sure that they keep all their software up to date. This removes the risk of bugs and hence removes potential vulnerabilities in the software.

Finally, DDos requires a technical solution to deal with the flood of requests.

DATA ARCHITECTURE

One of the key parts of data governance is data architecture. Architecture is all about how data is tracked and made use of in an organization. Why is this all important? Well, if you don't have a view of the exact data you have, how can you possibly govern it?

During this section we will talk about:

- Policies (again). It's a bit about data usage guidelines
- Data models - logical and physical data models
- Data catalogs - what have we got?
- Data lineage / provenance
- Third party system connectivity
- Certified data sources

Let's start with policies. We've already talked about them a lot, so I won't say much about them here. It's just to reinforce that we need a set of rules and policies in place that tell users what they can and can't do with the data that we store. As we know, humans are the weak link in data security, so leaving them in no doubt over what they can and can't do is vital.

Next, we are going to talk about data models. We have two types, logical and physical.

A physical data model shows the table schemas - that is, the column names, data types, domain restrictions (constraints), primary keys, foreign keys and relationships between tables.

The logical data model is a high level version of the physical model. It shows the fields involved and the relationships between tables, but it doesn't go into detail about the data types or constraints.

The physical model refers to the physical implementation of the tables. Like, name is a string with a restriction of 100 characters. The logical model just looks at how it’s going to hang together, without those implementation details.

Data catalogs are just like the old fashioned catalogs you might have received through the post for clothing. It lists out all the data available in your systems, along with data types and detailed field descriptions.

Field name	Data type	Length	Description

When you’re a company with tens of systems with strangely (and not very well thought out) field naming conventions, this sort of catalog is absolutely required, so you know what you have and where you can find it.

Next, we have data lineage, which is all about finding out what happened to your data. Where did it come from? What happened to it once it landed in the system? What ETL process ran on the data?

This is key functionality as it enables us as data managers to handle the data better. We know what has happened to it and where it came from, so we know that it has been collected in-line with the regulations and that nothing has happened to it to affect the accuracy.

Next, we need to define the rules around third party system connectivity. This is particularly important in the world of self service analytics. If you connect Tableau to your datasource, how do you make sure that the right people are seeing the right data - how are you going to make the connection? Is it a secure connection? All of this needs to be planned and mapped out before any action is taken.

We can also look to create certified data sources. This is where we create a datasource for users to consume. Again, this is key in the world of self service analytics, where we need to make sure that people are looking at accurate data and that we restrict access to the source data.

This lets users create their own insights without having access to data that we don’t want them to see. It’s an extra level of security, while also enabling additional flexibility.

KEY TAKEAWAYS

The key things we discussed in this book are:

1. Make sure you have policies in place for absolutely everything. Humans are the weak link in our data governance efforts and we need to make sure they know what they can and can't do and are held accountable.

2. Training is an absolute must. But more than that, we need to drive a data culture internally. We need to get people excited about the data and the possibilities it has for the business and we also need to hammer home the need to protect that valuable asset.

3. There are loads of types of attack and we understand that some are unavoidable but with the right data governance and data architecture, we can reduce the number of vulnerabilities we have and mitigate the impact for those that we can't avoid.

Technical solutions are always absolutely required, but in isolation they will not protect you from data breaches. We need the people (policies and procedures) to work in harmony with the technology to truly have a governed environment.