White Paper

White Paper by Bloor

Author **Philip Howard**

Publish date **November 2016**

Discovering data occurrence

> It is obvious that in order to protect sensitive data you need to be able to discover and identify it in the first place. Unfortunately, experience suggests that many companies do not do enough to identify and understand where sensitive information is held within their organisations.

Author **Philip Howard**

Executive summary

The original working title of this paper was with respect to discovering (and subsequently securing) information that is subject to data privacy and data protection regulations. However, personally identifiable information is not the only data that needs protection. Organisations possess multiple types of data that need securing. Examples would include formulae for proprietary products, salary information, pricing details that might be valuable to competitors, and so on. Thus we might have referred simply to discovering sensitive data and, indeed, the most common use cases for which the technologies discussed in this paper are with respect to sensitive information. Nevertheless, in researching the techniques that are available, we have concluded that some of these might have uses that go beyond sensitive data, hence the rather bland title we have adopted.

It is obvious that in order to protect sensitive data you need to be able to discover and identify it in the first place. Unfortunately, experience suggests that many companies do not do enough to identify and understand where sensitive information is held within their organisations. This is not entirely their fault: many vendors of IT tools claim to be able to discover this sort of data and it is certainly true that there are many suppliers that can identify social security or credit card numbers. However, there is a great deal more to personal information than nicely formatted sequences of numbers and letters and the truth is that most companies claiming to be able to identify sensitive data have tools that are inadequate for this purpose and have only limited capability. In the view of Bloor Research, additional techniques are required if you truly want to understand where sensitive information is stored across your organisation, and in this paper we will discuss the various techniques that are available for this purpose. As it turns out, one of these techniques is suitable for tracing any sort of data and not necessarily data that needs to be protected. Hence our comments in the previous paragraph.

Before discussing techniques, we will begin with a short recap on why it is important to be able to protect sensitive data.

> “...the truth is that most companies claiming to be able to identify sensitive data have tools that are inadequate for this purpose and have only limited capability.”

Protecting sensitive data

> “Had Target been operating in the EU and had GDPR been in force then, it could have been subject to a fine of approaching $3 billion.”

There are essentially two reasons for protecting sensitive data. One is reputation loss in the case of a data breach and the second is regulation and compliance. Arguably, the perceptions of customers are a third reason: a recent survey conducted by Experian and dataIQ found that 16% of consumers are happy to share their personal details *“if they trust the company involved”* but 49% would *“prefer not to share unless they have to”*. According to another survey, this time by DST and dataIQ, more than one fifth of consumers believe that their personal data should be deleted immediately and a similar number think that consent to use their information should only be valid for six months.

This deep-seated unease felt by customers is, of course, directly linked to the frequency of data breaches. The more they get publicised the more concerned the public feels. And the basic problem is not the data breach itself: it is the fact that whoever hacked the system can actually read and therefore leverage the information they have stolen. If data is appropriately secured, then it should not matter if there is a breach because that information cannot be used. However, this does imply that all relevant data is secured. And this is the issue – often it is only some of the data that is protected – and this means that sophisticated hackers may still be able to exploit the data they have stolen by exploiting correlations and patterns that allow identification of sensitive data.

The other side to data breaches are, of course, fines and court actions. Target for example, after its 2013 data breach, eventually settled claims of more than $100 million. You might think this would be sufficiently salutary that companies in general would ensure that they got their act together with respect to protecting sensitive data. However, when you bear in mind that Target has revenues (in 2015) of over $75 billion and that the estimated loss of direct sales revenue was 2.5% of one quarter’s earnings, then you can see that there is no great incentive to comply with data protection legislation. However, the EU is tackling this problem. The recently introduced GDPR (general data protection regulation), which will come into force for all companies collecting data about EU citizens in 2018, carries a maximum fine of 4% of global revenues for a data breach. Had Target been operating in the EU and had GDPR been in force then, it could have been subject to a fine of approaching $3 billion. Given the EU’s recent record (see its recent claim for $13 billion against Apple) it seems likely that the penalties for data breaches – at least those involving EU citizens – is likely to rise substantially. Further, we expect that other legislative bodies will take a similarly strict view in the future: so that data breaches – at least where personally identifiable information is exposed – will really hurt your bottom line.

It is worth expounding on GDPR a little more, both because of the increased emphasis on consent (more so than most others, such as HIPAA, PCI and so on) and because it extends the definition of what is considered personal information. For example, it is not currently the case that IP addresses need to be protected but they will have to be under GDPR.

From a consent perspective, there are several points. Firstly, the reason for collecting the data in the first place must be explained in such a way that is unambiguous and consent must be confirmed by affirmative action. In other words, opt-in not opt-out, and no hiding consent in the small print. It is notable that the DST/dataIQ survey found that only 15% of businesses track permissions company-wide, so this is something that will need to be addressed. Other notable aspects of GDPR are that young people below the age of consent cannot give consent (logical but not the way that Facebook currently view it), that consumers have the right to see the data held about them (which means that businesses will have to have a holistic view of all personal data) and to be able to correct that information, and that they may demand that their personal details are removed. For a more detailed discussion on GDPR and its implications for data management see (***The data management implications of GDPR***).

Discovery technologies

There are a number of technologies that can be used to discover sensitive and other data of interest, depending on the environment. We will discuss each of these in turn.

Data profiling for data quality

There are actually two distinct types of data profiling that are used for different purposes and we will discuss these separately. We will start with the most well-known form of this product type.

Data profiling has historically been used as both a precursor to data matching and cleansing and as an environment that provides on-going monitoring of data quality. In order to do that, profiling looks at each column in the database and compares its content with the metadata for that column. Thus, if the metadata says that this is a numeric column but some of the data entries have alphabetic content then those entries are in error (or, exceptionally, the metadata definition is in error). In addition, profiling tools can identify common formats within a column and compare each result to that common format. From this common format the software may be able to deduce, for example, that this is probably a column for email addresses or postal codes. Further, data profiling tools can identify instances where the @ symbol had been omitted and that this is therefore an invalid email address. In some other cases, the software may be able to establish whether a postcode, say, is valid or invalid. A third possible way in which data profiling tools work is to recognise data that is in a specific format, such as a social security or credit card number.

In addition to these base capabilities data profiling tools can identify relationships that exist within a database. For example, that a customer has an order, or multiple orders. Within a relational context it can follow primary-foreign key relationships and thus, once you have established a customer record you can also discover associated data that you might also need to protect. However, here we start to run into the first problem with data profiling tools, which is that not all of them are capable of following such relationships across databases. This is because, where there are relationships across data sources these are typically implicit rather than explicit and the relevant software tool has to therefore be able to infer relationships rather than follow them directly. Secondly, most data profiling tools were developed to work specifically with relational data sources and they typically (there are exceptions) have little or no ability to profile data in non-relational databases. And we don't simply mean NoSQL databases here, we also mean legacy systems such as IMS.

Unfortunately, there is also a problem of scale: it is not merely a question of this customer data in this database being related to data about the same customer in that database: often these things go in chains, with relationships spanning multiple data sources. And most data profiling tools haven't been built to support that sort of scale, especially when the number of databases in a large enterprise may stretch into the thousands or even tens of thousands. Note that a separate, but related problem, occurs when you have multiple copies or versions of the same database, which may be being used for back-up, development, testing, Q&A and so forth. Traditional data profiling tools have not been designed to cater to this proliferation of the same data.

To conclude this section, data profiling tools are useful for discovering specific patterns of data that can be precisely defined. They may also be used to infer relevant relationships of importance once an initial starting point (customer name, for example) has been established (either directly by the software or through human intervention). However, they may lack the ability to scale across very large and complex environments unless they have been specifically designed for that purpose.

Data profiling for testing

There is a class of data profiling tools – mostly provided by vendors in the test data management space – that profiles databases for a different purpose. Here the intention is to profile the database so that you can create a synthetic dataset that is representative of the real data but is, in fact, artificial. This can then be used for development and testing without impacting on your production database and without compromising the security of any sensitive data. One knock-on effect

> "…the intention is to profile the database so that you can create a synthetic dataset that is representative of the real data but is, in fact, artificial."

of the use of synthetic data is that any methods used to discover sensitive data (conventional data profiling or otherwise) need to be informed about or detect these data sources with synthetic data in them, so that these can be excluded from any analysis of the data landscape.

Semantics

It would be nice if you could automatically identify the contents of any database column (we are back to focusing on relational databases here) by reading and parsing the column name. Unfortunately, very few companies adhere to strict naming standards for database columns and, even where they do, they seldom use anything that is meaningful. It would be extremely useful if SAP, Oracle and other providers of ERP solutions used meaningful descriptions for all their columns. If, for example, every column containing a person's name had "name" somewhere in the column header. Of course this would also make the life of a hacker easier but it would also make protecting the date easier. Anyway, they don't do this.

However, this doesn't mean that semantics does not have any value here, even at a relatively crude level. For example, if you are searching a large database then simply having a table containing the one hundred most popular first names in any particular country, would surely be sufficient for you to infer any columns containing first names. Similar approaches can be used for countries, states, cities and so forth. In particular industry sectors a knowledge of a relevant ontology would be similarly useful. For example, in healthcare, treatments, operations, drugs and so forth might all be associated with sensitive patient information and a knowledge of the relevant vocabulary would be useful.

> "Unfortunately, very few companies adhere to strict naming standards for database columns and, even where they do, they seldom use anything that is meaningful."

Code introspection

Another method that is used in some instances to identify sensitive data is to introspect the applications and code that manipulates your data. Thus, for example, if you have an ETL (extract, transform and load) program that extracts personal data from your transactional database and loads it into a data warehouse, then it may be possible to introspect the ETL code to see what it is doing with the data it has extracted. This is important because the data may have been transformed during this process so any data profiling may not recognise that a relationship exists.

The problem with this approach is that there are a lot of ETL tools and there are a lot of APIs, there is a lot of middleware, and there are a lot of different programming languages, all of which can be used, not just to move data, but to transform it en route. Thus, as a general purpose approach this technique has limited applicability. Where it has been proved to be useful is for introspecting database stored procedures. This only requires the ability to understand SQL and relevant SQL variants. However, there is again a question of scale: it is one thing introspecting the stored procedures used in one or a few databases, but doing it at scale is another thing entirely.

Data catalogues

We should mention that there a number of vendors that have developed data cataloguing techniques for use in conjunction with data lakes. Basically, these work by identifying the metadata that describes the data poured into a data lake and then create a searchable catalogue of that metadata. While limited in scope these products may be useful in identifying sensitive data held in data lakes that needs to be protected. For example, under GDPR twitter handles potentially represent sensitive data, which could impact on sentiment analysis.

Discovering associations

There is one further technique that we are aware of, not yet in production, but under development, that is worth mentioning. Suppose that you on-board a new customer. This data flows through various systems and makes updates to various fields in various databases. You want to discover all of these associated locations. In general, these updates to other systems will occur within a certain time period. Thus, if you monitor relevant databases and see that whenever you on-board a new account in database x then within five minutes there is an update to a table in database y, then you can reasonably infer that these are related activities. You might want to have a data steward confirm this inference but this technique would allow you to discover relationships that might not otherwise be apparent. In effect, what this approach does – like code introspection – is to discover the non-obvious. The big advantage of this method is that it is not limited to stored procedures.

Sensitivity analysis

All of the previously discussed approaches have problems with scale. Unless specifically designed otherwise – and almost all products are not – they cannot cope with very large ecosystems. However, that is not the only problem: none of these techniques, taken in isolation, is sufficient to find all the data you are potentially interested in. Combining techniques: say data profiling with semantics plus code introspection might be sufficient if there was not the problem of scale. However, there is a further issue. In large enterprises with many databases it is not simply a question of identifying all of your sensitive data and then masking or encrypting it. Doing this is a potentially mammoth task. You therefore need to understand how sensitive different data elements are, so that you can establish priorities. In other words, you need a way to classify sensitive data. To take a simple example, IP addresses under GDPR are considered sensitive but that is not the case in other jurisdictions. Thus protecting the IP addresses of EU citizens would have higher priority than those of nationals of other countries. Similarly, credit card and social security numbers would have very high priority whereas gender information might be less important. Note that this will be industry dependent: gender might be considered very important in healthcare but less so in financial services.

Sensitivity analysis should be automated. There will be an initial set-up process through which you define the priority to be associated with relevant data elements and this will need to include parameters to allow for different geographical and other effects, but once set-up this process should run automatically. Note that support for semantics will be important in providing this automation, because you want the software to recognise for itself that this is a first name, a surname, an address field and so on.

> “You therefore need to understand how sensitive different data elements are, so that you can establish priorities.”

Conclusion

It would be easy to conclude that it is a simple matter to identify sensitive data. The truth is that this is not the case. Far too often, vendors glibly claim that of course they can do this when the fact is that they are only partially successful, at best. Of course, not all vendors are equally culpable of exaggerating their capabilities but we have encountered a degree of complacency that is particularly obvious from some of the suppliers of data quality solutions: who seem to think that traditional data profiling can do everything. In our view, this is not the case. In this paper, we have tried to highlight the real issues and the technologies that are available, or becoming available, in discovering where sensitive data exists. Companies that are serious about meeting compliance and governance standards with respect to privacy – and sensitive data more generally – need to consider the issues raised here and match these against potential providers.

> “...we have encountered a degree of complacency that is particularly obvious from some of the suppliers of data quality solutions.”

FURTHER INFORMATION
Further information is available from
www.BloorResearch.com/update/2309

About the author

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter 'noise' and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channels.

Founded in 1989, we have spent 25 years distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

Copyright and disclaimer

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**