aircloak

Data Anonymisation

What it is and Why it Matters

We find ourselves today in the midst of a data revolution. Exciting new technologies like artificial intelligence, the internet of things, and self-driving cars are all made possible by the new science of data.

A good analogy is that data is to this century as oil was to the last. Today, data drives growth and change, opening up new avenues of risk and innovation. Everything, from trains to toasters, now has the potential to be a source of valuable, analysable data.

Understandably, this trend shows no signs of slowing down. In fact, it's expected to experience explosive growth in the coming years, with the big data industry reaching $66.8 billion by 2021.[i]

The Rise of Data Protection

This new data gold rush has brought with it increased concern around data security. High profile corporate data breaches have made customers wary of what is happening to the vast oceans of their personal data now being collected. And after Edward Snowden's revelations about the post 9/11 erosions of citizens' rights, the average citizen is no longer willing to stand idly by as corporations profit from their data while putting it at risk.

The General Data Protection Regulation

The General Data Protection Regulation (or GDPR) is a new piece of legislation passed by the European Union to protect the personal data rights of EU citizens. As of May 25, 2018, many new data protections will come into force. For example, individuals will be given the right to access any of their own personal data being held by companies. They will also gain the so-called "right to be forgotten," which is the right to have their personal data destroyed upon request.

Under the GDPR, companies will need to implement "data protection by design and by default."[ii] They will also be required to promptly notify the Data Protection Authority and any affected individuals in the event of a data breach. And consent to use personal data must be freely and specifically given rather than buried in a sea of complex legal jargon.

One important thing to be aware of is the breadth of the GDPR's jurisdiction. The new regulation applies to all companies collecting, storing, or using the personal data of EU citizens, regardless of where they currently reside.

What happens if you fail to comply? The fines are no laughing matter: 20 million Euros or 4% of annual global turnover, whichever is higher. And the fun doesn't end there. The EU will also have the power to ban non-compliant organizations from trading with any nation that has adopted the GDPR into national law.[iii] Not a bad motivation to update your organization's data security policies.

Is the GDPR Bad for Business?

It doesn't have to be. For one thing, it makes EU data privacy legislation clearer and more consistent. In addition, with all the public focus on personal data security, there is a real opportunity for companies to differentiate themselves through superior data privacy practices.

Smart companies will use the GDPR as the perfect time to revisit and improve their data privacy programs.

"The GDPR comes as a welcome piece of legislation, and in many ways the reform creates a strong and comprehensive set of rules that need to be applied in order to sufficiently protect data," says Phil Bindley, CTO of The Bunker. "In doing so, this creates more trust in the digital environment and means that a privacy-focused approach can work in conjunction with the use and analysis of data."[iv]

Loss of Reputation from Data Breaches

If your organization suffers a data breach, you may have more to worry about than massive regulatory fines. In a recent survey conducted by the Ponemon Institute, 75% of companies said their reputation was a key asset but fewer than half said they thought that reputation could withstand a negative event such as a data breach.[v]

Additional research bears this out. According to data from Semafone, the overwhelming majority of people would stop doing business with a company that had suffered a breach. As Tim Critchley of Semafone put it, these numbers prove "what we should already know—that the reputational damage suffered by companies who fail to protect personal data can translate directly into loss of business."[vi]

Data Anonymisation vs Pseudonymisation (and Why it Matters)

With all of the new data analytics and increased scrutiny around data privacy, tomorrow's winners will be those who figure out how to leverage insights without compromising privacy.

One of the most powerful tools for this quest is data anonymisation. Data anonymisation is so important because it allows you to leverage your data for marketing insights without jeopardizing individual data privacy.

Data protection laws exist to protect the personal identity of people whom the data describes. If the data subject is not identifiable in any way, data protection law does not apply. That is, when it becomes impossible to connect individuals to the data, people controlling and processing the sensitive data are not restricted in data use or sharing. On the other hand, an identifiable data subject can result in legal consequences including damage claims, loss of reputation, and fines or penalties.

The purpose of data anonymisation is privacy protection. It involves the modification of data sets so that no personally identifiable information remains. As a result, data can be used and transferred without individuals' identities being disclosed unintentionally.

Pseudonymisation differs in that personally identifying fields in the data are replaced by pseudonyms such as random numbers. Simply replacing these fields, however, does not eliminate the possibility of identifying individuals in the data. In fact, a study of human mobility data revealed that four spatiotemporal points are enough to uniquely identify 95% of the individuals in a dataset, even if the granularity is low and you have no additional information about a user. [vii]

For a mobile call records dataset like this, pseudonymisation might involve replacing people's names and phone numbers with random (although not necessarily arbitrary) numbers.

However, this is no protection against what's known as an inference attack. For example, if the analyst knew of 2 or 3 calls that a friend made at certain times from certain places they could be able to (re)identify their friend's records from the pseudonymised data.

True data anonymisation is difficult to achieve, and many data controllers fail to do so properly and completely. Proper anonymisation often destroys the value of data to the point where the whole exercise becomes moot. This happens because when you anonymise a dataset before analysing it, you have to anonymise it across all dimensions that are identifying (or potentially identifying) in combination with other attributes or in combination with external knowledge. Understanding what could be identifying (beyond the obvious things like a social security number of phone number) is difficult and usually underestimated.

In fact, most companies use the weaker pseudonymisation techniques to protect personal data, which also means many companies will be constrained by data privacy laws and subject to penalties when the GDPR comes into effect.

The EU has developed a list of criteria for anonymisation with their WP29 Opinion on Anonymisation Techniques. First of all, the WP29 makes it clear that pseudonymisation is not strong enough to protect personal data. Pseudonymisation permits data controllers to handle their data more liberally, but it does not abolish all risks due to the possibility of re-identification. Crucially, as a result, pseudonymous data is still subject to privacy regulations under the GDPR.

Overall, the WP29 is intent on educating organizations about proper anonymisation techniques and are calling for certification so that companies follow their requirements. Data protection authorities (DPAs) can carry out this certification. Yet, no clear guidance currently exists on how data anonymisation techniques should be certified.

Because there are no specific guidelines on how to ensure anonymisation, DPAs are certifying data anonymisation practices on an ad hoc basis. Instead of following specific guides, organizations are making certifications based on their own interpretations of the GDPR's loose criteria. There is no general anonymisation certification program that companies can follow before being certified. It is for instance not clear what happens legally

when a certified anonymisation process is later discovered to be weak or performed by someone who is unqualified. It is difficult to tell who is legally liable in such situations, and non-compliance can result in steep penalties after the GDPR comes into effect.

There are a number of traditional data anonymisation techniques, like K-anonymity and differential privacy, that are simple and provide strong anonymity but unfortunately destroy data utility.

To retain data utility, data controllers typically choose from a complex variety of mechanisms that provide some anonymity but may not protect against re-identification in many cases. These include rounding, cell swapping, outlier removal, aggregation, sampling, and other techniques. Getting this right requires substantial expertise both on the part of the data controller and the DPA. And again, if done right, this process typically destroys the utility of the data

Diffix: A new approach to data anonymisation

Aircloak, in research partnership with the Max Planck Institute for Software

Systems, has put many years into researching Privacy-Enhancing Technologies (PETs) and have introduced a new approach to anonymising data sets called Diffix.

Diffix is the base of Aircloak's flagship product, Insights™, a solution for these privacy regulation problems. Aircloak Insights provides strong anonymity and good utility for a wide range of use cases and requires no special expertise to setup and configure. Aircloak Insights maintains the quality of the data set for analysis but also achieves a strong level of anonymity.

Because Aircloak Insights works "out-of-the-box" for a wide range of use cases, once a Data Protection Authority has certified data anonymised by Insights, any new use case requires little to no additional evaluation. CNIL, the French national data protection authority, has evaluated Diffix against the GDPR anonymity criteria, and have stated that Aircloak delivers GDPR-level anonymity.

As a result, data that is anonymised by Aircloak's solution will not be constrained by privacy laws or be the subject of penalties when the GDPR comes into effect.

Large European Banking Client Reduced Complexity of GDPR Compliance Processes with Aircloak

The client had traditionally been a provider of consumer credit products. They wanted to find ways to acquire high-value customers, better understand their existing customer base, and offer new products. This was their inspiration behind the creation of a mobile application that allowed end-users to aggregate their bank account information and financial transactions. In return, it provided valuable financial insights and recommendations.

The client relies on continuous detailed customer analytics to better understand their customers' financial situations. This knowledge makes it possible to deliver the best recommendations and advice to customers and improve the usability and functionality of their app without creating unnecessary features.

While enhancing user experience is a business priority for the client, so is guaranteeing the privacy of their customer data. It is necessary that the company ensures that nobody - not even their own analysts - can access individual user data. Therefore, all aggregate analysis of customer data must be carried out with the highest safety and anonymity possible.

Prior to using Aircloak, they achieved safe aggregate data analysis using a classical approach.

Whenever a new type of exploratory analysis was considered, an approval process had to be followed involving various internal stakeholders and privacy professionals. This process was complex and could take weeks to complete.

Following the approval process, the client would provide a snapshot of their data set to a third-party expert organization who would anonymise it using traditional data-masking techniques. While safe, this process resulted in additional delays and a severe reduction in data quality. At times, the resulting data quality would be so low that a new analysis approach would need to be devised, beginning the process all over again. Additionally, whenever new data became available, the anonymisation procedure had to be repeated.

Pain Point: A Slow and Complex Anonymisation Process

A key problem for the client was the fact that the quality of free form text information suffers greatly from anonymisation. This is because only white-listed terms were allowed through while everything else is immediately removed or made illegible. Changes to the white list required new approvals which cause additional delays.

The client's approach, although compliant, was far too slow and complex. By working with third-party data processors, they introduced risks to the system and were left with inaccurate data that impaired analysis accuracy.

The Client needed a new solution to analyse customer financial transaction data that:

- Satisfied the highest level of privacy protection available and was fully compliant with current and future privacy regulatory requirements (GDPR)
- Was simpler and faster to execute
- Did not introduce dependencies on external data processors
- Did not destroy valuable information in their data

Aircloak's Fast and Simple Solution

Aircloak Insights takes a fundamentally different approach to securing the privacy of databases. Architecturally, Aircloak Insights acts as a proxy between a database and an analyst by querying the data through the product's query editor, by using Aircloak Insights' API, or by using Tableau Desktop – a software designed to help people see and understand their data.

Instead of anonymising or pseudonymising the source data to the point of potentially becoming unusable, analytics are done on the unprocessed data. The analytics output is instantly and automatically anonymised as queries are run, thereby guaranteeing the strongest level protection of a client's private customer data by design. The key to this process is the fact that the analytics and anonymization steps are one and the same. If anonymization is done before analytics, the analytics suffers, if it's done after then anonymization cannot be ensured. Because Aircloak has the technology to do it all at the same time, it provides both strong anonymity and good data quality.

And because Aircloak Insights automatically ensures that no personal information ever leaks, our client has been able to greatly optimize their existing process of analysing private customer data in the following ways:

- Any approved internal stakeholder can now run arbitrary queries over the raw data.
- Zero dependence on a third-party processor eliminates a source of risk, cost and complexity.
- Minimization of the delays for the entire process.
- The client can be sure that they are compliant with all current and future data protection regulations.

Client Benefits of Aircloak Insights

Now able to use the full transaction details for analysis in a secure and compliant way, our banking client can better understand and segment its customer. This allows them to make better recommendations regarding the customer's financial situation and potentially relevant products.

Aircloak Insights provided the client with a solution that ensures compliance with privacy and data protection regulations. In particular, a European data protection authority has verified that Aircloak satisfies the anonymisation requirements of the GDPR.

Instant Privacy Compliance: Share Insights Not Data

Aircloak's first-in-class real-time database anonymization solution provides instant privacy compliance and enables high-quality analytics for any data set and any use case. Aircloak has been evaluated and approved for GDPR-level anonymity for all use cases by CNIL, the French data protection authority.

Aircloak unlocks sensitive data for analysis, sharing, and monetization. Aircloak's strong anonymisation satisfies the world's strictest anonymity laws, slashing compliance costs and all but eliminating data sharing risks.

Aircloak's patented breakthrough technology is the first anonymization solution to provide high-quality analytics, unprecedented ease-of-configuration, and strong anonymity. Prior approaches that use pseudonymisation provide good analytics but weak anonymization, falling well short of European standards. Prior forms of strong anonymization, like K-anonymity, tend to destroy the data in all but the simplest use cases.

Aircloak's ease-of-configuration far exceeds all previous solutions. There is no need to flag or remove Personally Identifiable Information (PII), pseudo-identifiers, or sensitive data. The existing primary-use database is not modified in any way, and Aircloak handles all data types including unstructured text. Aircloak is installed in front of the database, and relays queries and answers between the analyst and the database.

Aircloak offers analysts a rich explorative SQL database interface. Analysts submit SQL queries and interact with the existing database to extract the requested data. Both queries and answers are dynamically modified by Aircloak to ensure anonymity while still providing high accuracy.

The analyst has the look and feel of querying the database directly. The analyst can specify tables and columns, join tables, set ranges, run standard or custom SQL aggregate, math, and string functions, and much more. The analyst receives answers at the speed of SQL, thus

allowing the same interactive and explorative analytics the analyst would expect from native access to the database.

Aircloak enables new use cases; for example, improving internal business intelligence, external data monetization, and reducing the cost of mandatory reporting. Internally, divisions within large corporations can use Aircloak to share customer data with other divisions or out-sourced analysts. Organizations with valuable data, such as health, location, or financial data, can instantly and safely monetize the data regardless of its format. Aircloak gives organizations full visibility into how their data is being used, leading to better billing models.

Contact us today and let us show you how we can enable secondary use, allowing you to share insights, not data.

Aircloak GmbH
Gormannstrasse 14
10119 Berlin
Germany solutions@aircloak.com

---

[i] http://www.information-age.com/big-data-vs-privacy-big-balancing-act-123461795/

[ii] https://gdpr-info.eu/art-25-gdpr/

[iii] https://www.red-gate.com/simple-talk/opinion/opinion-pieces/questions-gdpr-shy-ask/

[iv] http://www.predictiveanalyticsworld.com/patimes/big-data-vs-privacy-balancing-act/7975/

[v] https://www.experian.com/assets/data-breach/white-papers/reputation-study.pdf

[vi] https://www.csoonline.com/article/3019283/data-breach/does-a-data-breach-really-affect-your-firm-s-reputation.html

[vii] https://www.nature.com/articles/srep01376