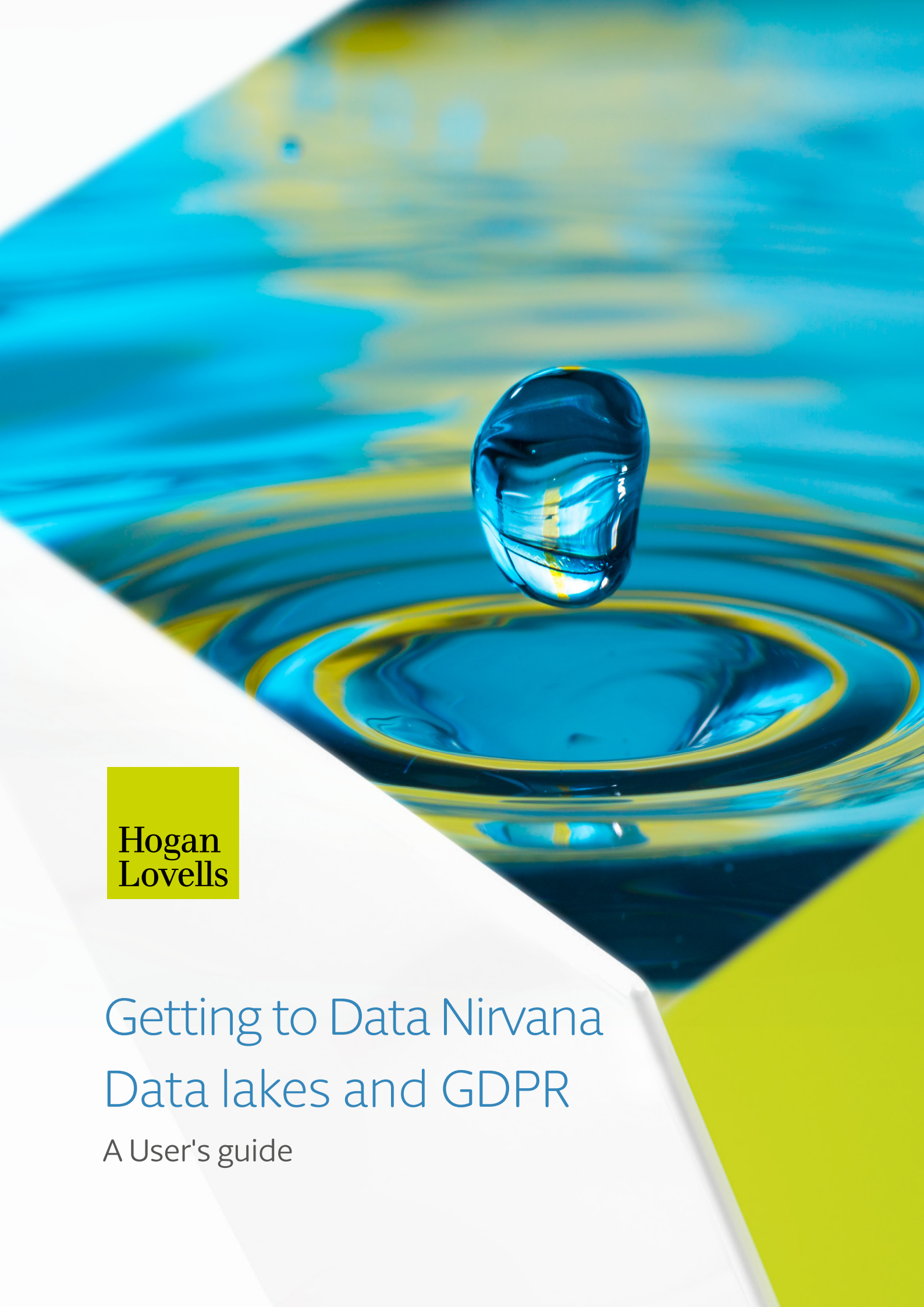
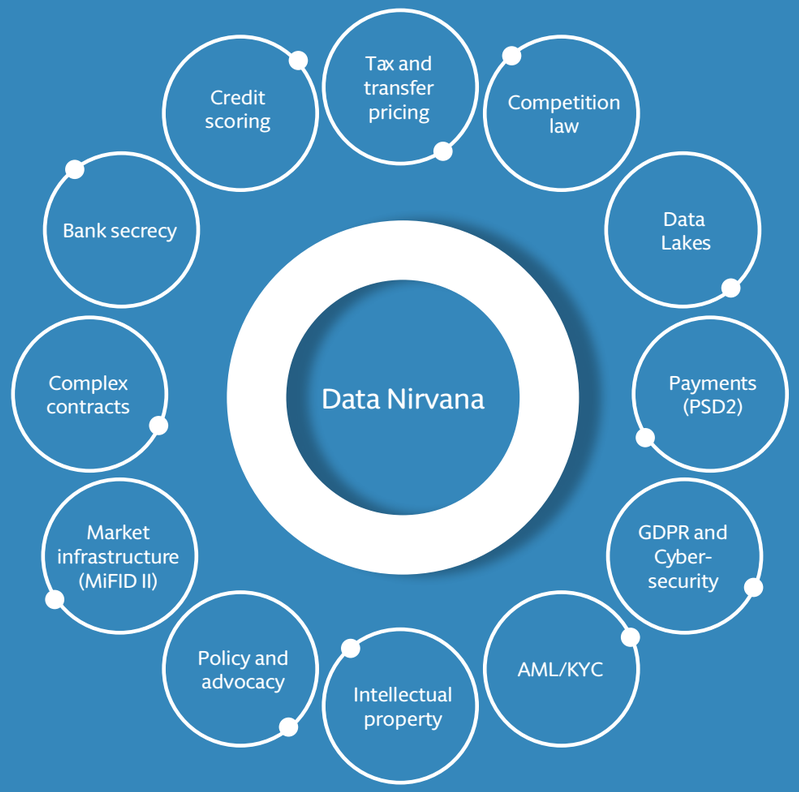


Hogan Lovells

Getting to Data Nirvana
Data lakes and GDPR

A User's guide



Winston Maxwell,
Partner,
Paris

Harriet Pearson,
Partner,
Washington, D.C.

John Salmon,
Partner,
London

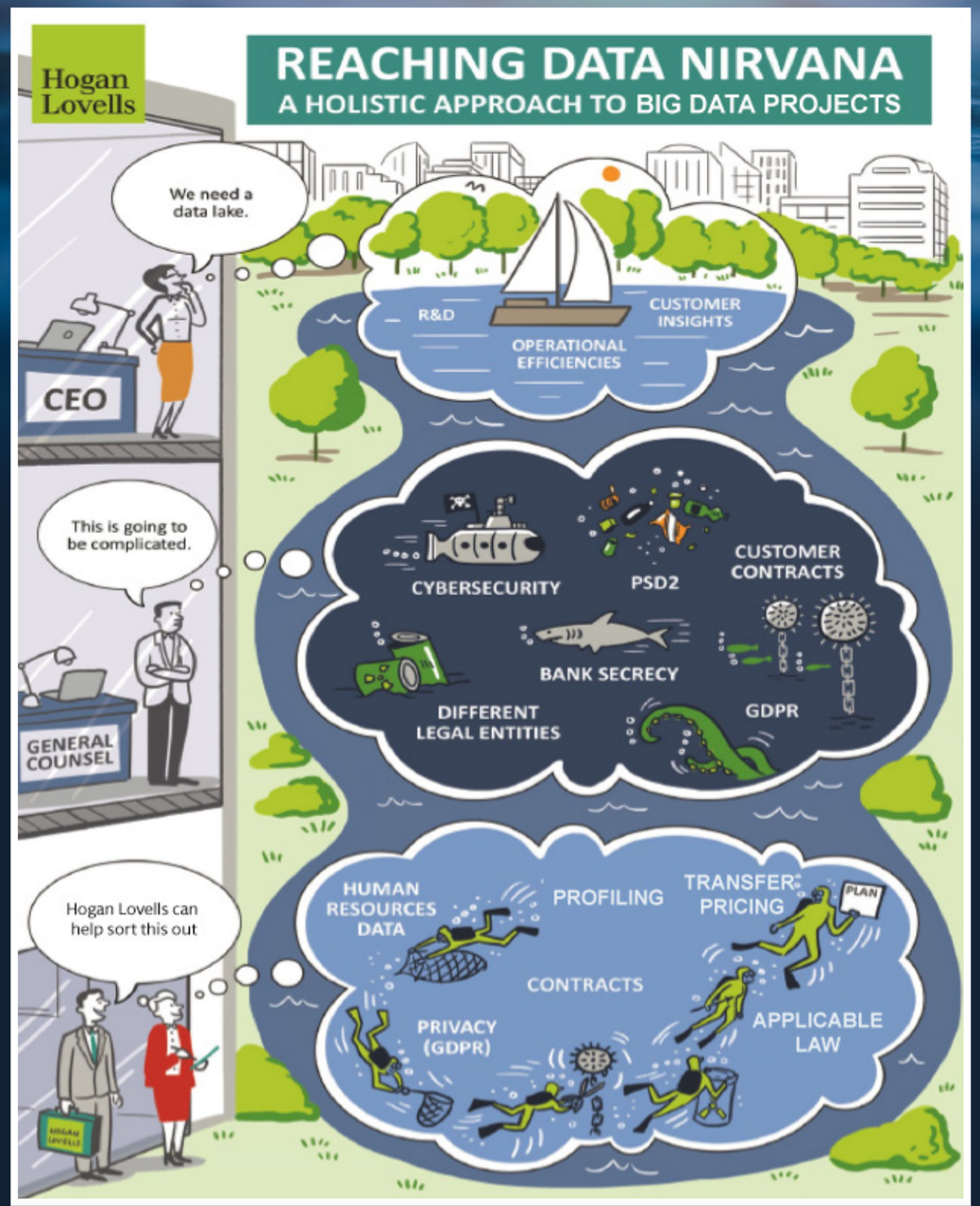
Eduardo Ustaran,
Partner,
London

Contents

Introduction: what's a data lake?

A data lake is an infrastructure that permits different data sets from within a group to be combined and analysed together.

To analyse a data lake under GDPR, it is helpful to think of a data lake in two phases: first at the infrastructure phase, and second at the applications phase.

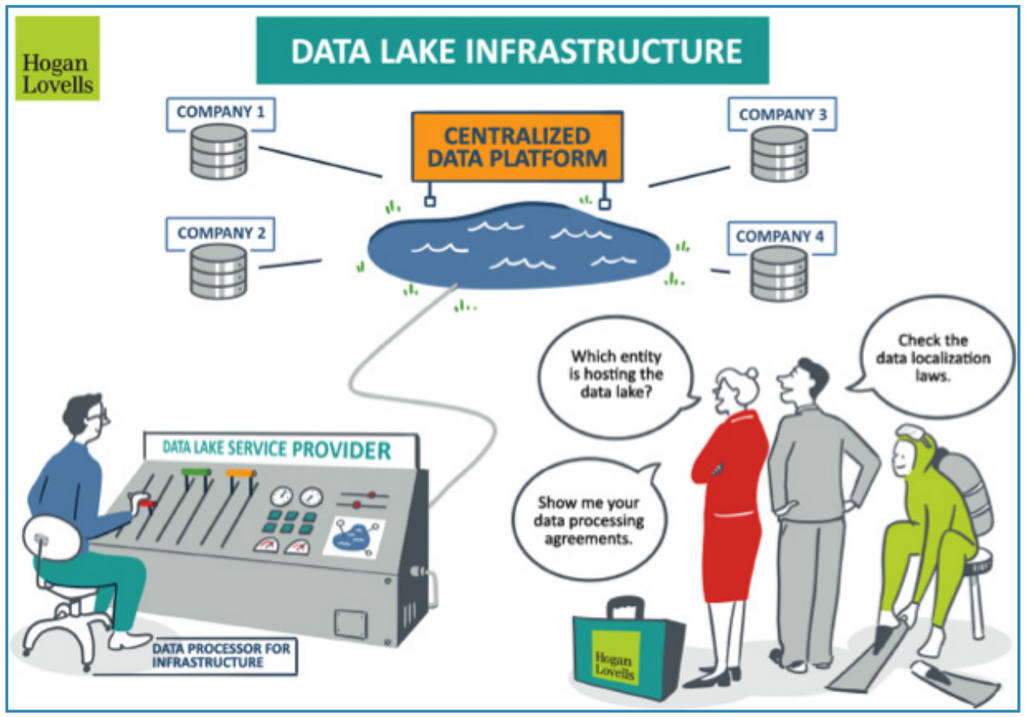


1. Data Lake infrastructure

For GDPR purposes, creation of the data lake infrastructure is similar to implementing a cloud infrastructure for group data. Upon creation of the data lake infrastructure, nothing is being done with the data other than moving a copy to a central repository, the so-called "data lake". Nevertheless, storing a copy of the data in a central repository is a form of processing under the GDPR. The entity hosting the data lake infrastructure will be considered the data processor, and the affiliates whose data are being stored will be considered the data controllers.

1.1 Identify the entity that is hosting the data lake.

The first step in any data lake infrastructure project will be to determine which entity in the group is providing the data lake infrastructure. We will call this entity the "data lake service provider". At this stage of the project, the data lake service provider will be considered a pure data processor, executing the instructions of each affiliate that has stored a copy of data in the data lake. The data lake service provider may the parent company, or it may be a dedicated IT affiliate that provides IT services for the group. The location of the data lake service provider may have an impact on applicable law. (see paragraph xx below).



1.2 Implement an intragroup data processing agreement.

At the infrastructure phase, the data lake service provider must enter into a data processing agreement with the affiliates contributing data, i.e. the data controllers. Like any data processing agreement, the data lake infrastructure agreement will need to comply with the conditions of Article 28 of the GDPR. This agreement should stipulate that the data lake service provider is holding the data in accordance with the data controller's instructions and should impose on the data hosting provider the obligations of security that flow from the GDPR. At this point in the project, there will be no specific purpose for which the affiliates' data are being centralised. The purposes of the data processing will be revealed later, at the applications stage.

The data processing agreement will state that the data is being centralised in order to facilitate future data analytics should the data controllers so instruct. However, until the data controller gives its express instructions, the data lake service provider is only allowed to store the data on behalf of the data controllers, which continue to exercise all rights over the data. The data processing agreement will also stipulate that the entire process is reversible, i.e. that upon instructions from the relevant affiliate, the data lake service provider will delete or return the data to the affiliate.

“
This agreement should stipulate that the data lake service provider is holding the data in accordance with the data controller's instructions and should impose on the data hosting provider the obligations of security that flow from the GDPR.
”

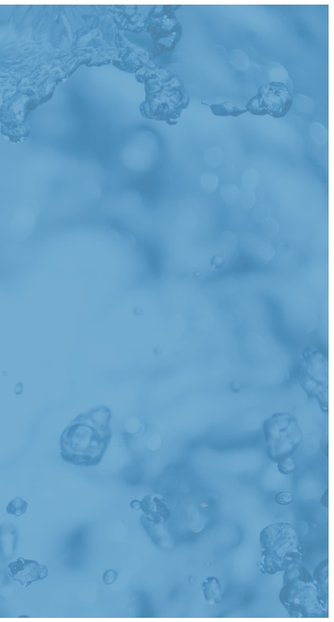
1.3 Check data localisation rules

Moving a copy of data to a central platform may raise other regulatory issues, particularly in countries that have data localisation requirements, such as Russia, China and Indonesia. Some countries do not allow certain kinds of data to be exported. This aspect would have to be verified before data are transferred to the data lake service provider. Similarly, if the data lake service provider is located outside the EU, the export of data from the EU would have to be covered by standard contractual clauses or some other adequacy mechanism.

1.4 Data protection impact assessment

A technical risk assessment conducted by the data lake hosting provider should be made available to each of the affiliates that are contributing data to the data lake. This technical risk assessment will not yet address the risks associated with each data lake application, but will be solely limited to the data lake's technical security measures, i.e. do any of the pipes leak? This technical risk assessment will form the baseline against which local affiliates can judge the adequacy of security measures put into place by the data lake service provider. At a later stage, an impact assessment may be necessary with respect to each of the applications that use the data lake infrastructure. But at the stage of the creation of the data lake infrastructure itself, the data lake service provider should conduct a technical and contractual assessment of the risks associated with unlawful access or loss of the data. The technical assessment would help establish that the data lake service provider has deployed appropriate technical and organisational measures to protect the data against loss.

“
The technical assessment would help establish that the data lake service provider has deployed appropriate technical and organisational measures to protect the data against loss.
”



1.5 Data lake governance committee

The data lake service provider within the group should create a data lake governance committee that can address questions and complaints of the affiliates that are contributing data to the data lake. Each affiliate in the group should be free to raise concerns about the data lake infrastructure, the content of the data processing and transfer agreements, or even the adequacy of the technical measures put in place. These concerns should be addressed by a centralised data governance committee and the responses to the concerns should be documented to ensure accountability. The same governance structure will naturally apply with regard to the subsequent data lake applications, because each application may give rise to separate concerns on behalf of the affiliates.

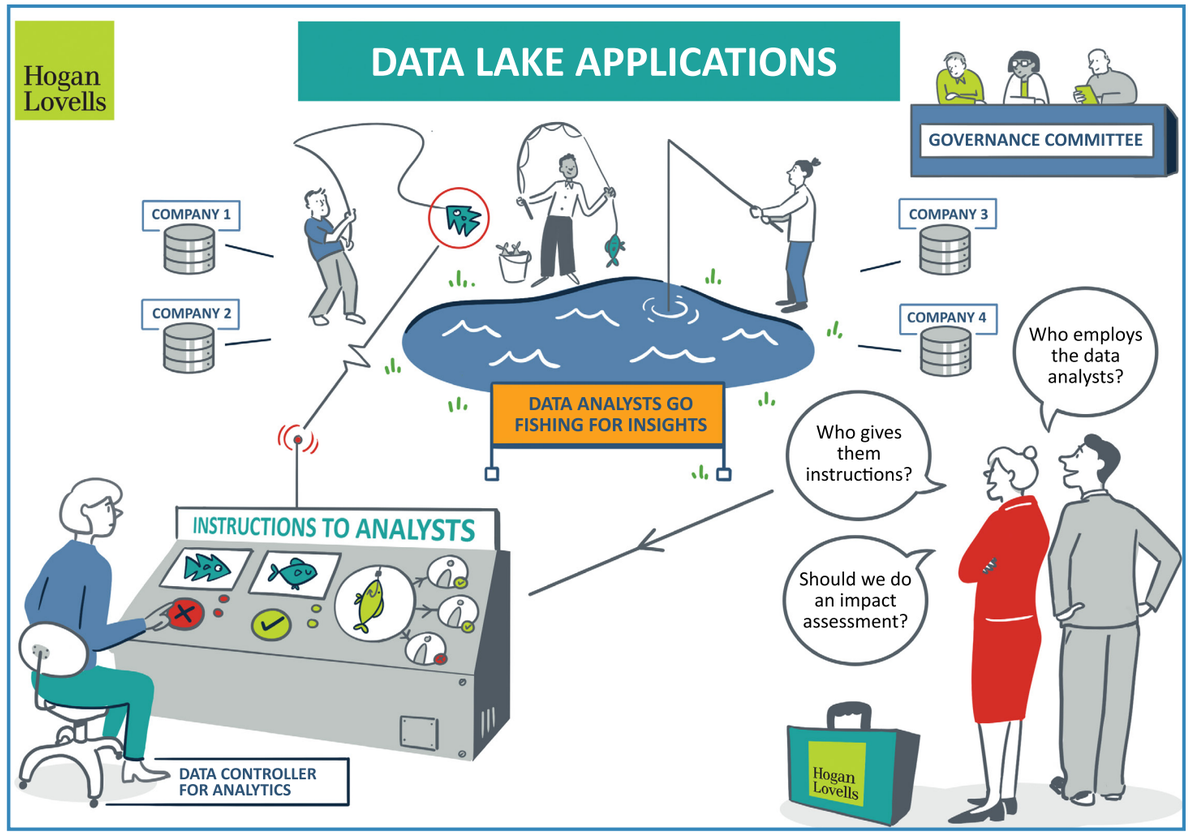
Each of the affiliates is accountable in the first instance for what is done with its data. It is therefore critical that the affiliates be able to make their own judgment regarding the adequacy of the measures put in place with regard to the data lake, and that the affiliate's concerns be addressed by a collegial body that represents the legal and compliance functions as well as the technical IT functions of the group. The governance committee would also permit modifications to the data processing agreement to be proposed and implemented.

2. Data Lake applications

2.1 Data lake service provider becomes data controller

Things become more complicated at the application stage, because the data lake service provider will also start analysing the data contained in the data lake. The first questions to address are: who is doing the data analytics and why? At the time the data lake is created, the data lake service provider will only have the role of data processor, following the instructions of the affiliate who provided the data. But as applications go live, the data lake service provider may transform itself into a data controller. In most data lake use cases, the analytics will be performed at the request of the parent entity, either for regulatory reporting or for purposes of developing detailed customer analytics.

But what then becomes the role of the relevant affiliates who contributed the data in the first place? Those affiliates remain data controllers, and remain responsible for the use of the data in the data lake. The affiliate may agree to allow the parent company to use the data in the data lake for a particular purpose as a data controller, but this decision belongs to the affiliate and must be documented.



2.2 Instructions from each affiliate as (original) data controller

In any given data lake configuration there are likely to be dozens of affiliates that are data controllers, each of which has transferred a copy of its data to the data lake service provider. Typically, data analytics will consist of analysing all the data from the dozens of affiliates and making correlations between the various data in order to create insights and value. The insights derived from the data analytics may benefit each affiliate as well as the parent company. The instructions from each affiliate to the data lake service provider may look as follows:

"In your role as data processor, please analyse my data together with the data of any other affiliates you have available in order to derive insights that may be valuable to me with regard to my customers or products."

The instructions may continue:

"As data processor, you are also authorised to provide similar insights to other affiliates in the group, including insights derived from data that I (the original data controller) provided."

Alternatively, the original data controller may acknowledge that the data lake service provider is also a data controller:

"In addition to storing my data in your role as data processor, you may apply data analytics for purposes determined by you at a group level, as data controller."

Whatever option is chosen, the original data controller should document its instructions, and make sure that the original data processing and data transfer agreements are updated.

2.3 Mapping value transfers from data lake applications.

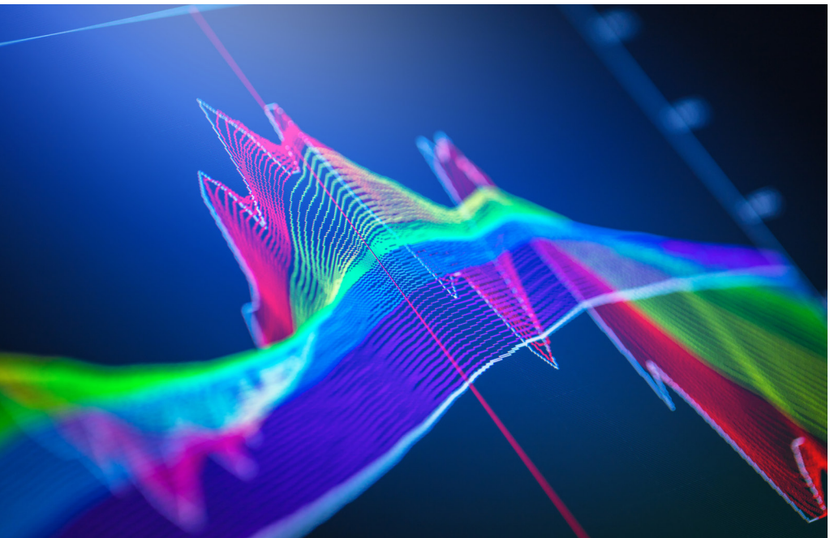
When data lake applications start, one of the first steps is to map out the transfers of value that will occur as a result of the data analysis, both for the affiliate that provided the data in the first place and the other affiliates who may benefit from the data analytics. The process of mapping out the value transfers is essential in order to establish appropriate transfer pricing documentation and also to help keep the roles of the affiliates and the data lake service provider clearly separated. Each affiliate should be getting value in exchange for the data that it is making available to the data lake service provider and the insights that are shared with other affiliates in the group. The value mapping exercise is not likely to be fundamentally different for each kind of application. Generally, the insights derived from a data lake will create value for each of the affiliates participating in the data lake but also for the parent company who is in charge of developing new products at a group level and who is in charge for regulatory reporting. Consequently, the value transfers for each kind of application will generally be similar.

“Generally, the insights derived from a data lake will create value for each of the affiliates participating in the data lake but also for the parent company who is in charge of developing new products at a group level and who is in charge for regulatory reporting. Consequently, the value transfers for each kind of application will generally be similar.”

2.4 Data lakes and applicable law

From a GDPR standpoint, each data application will have to be analysed on its own. An important threshold question relates to applicable law. For any data lake project, there will likely be a combination of personal data relating to EU residents and data relating to non-EU residents. The EU rules on data analytics are more restrictive than rules in many other parts of the world. For a data lake that consists of partly EU data and partly non-EU data, what law applies? Under the GDPR there are two main tests for applicable law. The first is whether the data controller is established in the EU. The second is whether the data controller, without being established in the EU, is nevertheless offering products or services to EU residents. In a data lake project, the data controller will typically be each individual affiliate. For example, in a retail banking scenario, the data controller will be the local retail affiliate holding the banking licence. Where that local affiliate is located outside the EU the GDPR will not generally apply because the data controller is located outside the EU and the processing does not relate to EU residents.

An important threshold task will be to map out applicable law in connection with each aspect of the data lake project. Different laws may apply to different aspects of the project. For example, if a U.S.-based affiliate transfers a copy of data to a EU-based data lake service provider, the rules determining what the U.S.-based data controller can do with the data will be governed by U.S. federal and state law. The data processor located in the EU will be subject to the GDPR in connection with its processing of data on behalf of the non-EU data controller. However the GDPR's application will be limited to the data lake provider's role as data processor. The service provider will be required to apply appropriate technical and organisational measures to ensure the security of data. However the GDPR will not affect the rights of the U.S.-based data controller with respect to its own data under U.S. law.



2.5 Anonymisation

When considering each kind of data lake application, one of the first questions relates to anonymisation. Data analytics performed on anonymised data fall outside the scope of data protection legislation. The problem is that many data analytics projects cannot usefully work on anonymised data. The standard for anonymisation under EU law is high. Under the European Court of Justice case law in Breyer[1], data remains personal as long as it is possible to identify indirectly an individual without disproportionate effort in terms of time, cost and manpower. The Breyer case involved a dynamic IP address, and the court held that because legal mechanisms exist to trace the IP address to a given subscriber, it was reasonably possible for the data to be traced to a given individual.

Another question is what does it mean to "identify" an individual. The highest French administrative court, the Conseil d'Etat, addressed this issue in connection with data in which the mobile device code that would permit identification of a mobile terminal was replaced by a hash code.[2] The technique called "salted hashing" prevents any person from reverse engineering the hash code in order to find the actual device number. The French Supreme Administrative Court found that this was not sufficient, because the purpose of the hash code was to single out a given user so that the user could be traced over time in order to observe the user's behaviour. The purpose of the data processing was to single out particular individuals and to study the patterns that emerge from those individuals' behaviour. According to the French Supreme Administrative Court, that is sufficient in itself to conclude that we are in the presence of "personal data". It is not necessary to be able to identify the person by name or address. It is sufficient that an individual is singled out for study. The European Court of Justice has not directly addressed this question, i.e. whether singling out an individual is sufficient to constitute personal data. However, recital 26 of the GDPR uses the words "singling out", suggesting that the position of the French Supreme Administrative Court may also apply to the question of personal data under the GDPR.

Regulators may also give some flexibility in borderline cases. For example, some data protection authorities in Germany have acknowledged that in the context of clinical data, a system of double pseudonymisation may be sufficient to reach the threshold of "anonymisation" under German law. There is a fine line between sophisticated pseudonymisation techniques and complete anonymisation. However from a practical standpoint, data controllers focussing on data lake applications should approach the problem as if they were in the presence of pseudonymised personal data. The burden of proof should be on the data analysts to prove that the data are fully anonymised under EU standards.

2.6 Additional purposes that are "compatible"

The practical consequence of data not being anonymised is that certain projects may not be possible because they relate to a purpose that is incompatible with the original purpose for which the data were collected. The relevant standard is set forth in Article 6-4 of the GDPR, which addresses the situation of processing for new purposes other than the purposes for which the data were originally collected. Article 89 of the GDPR allows member states to make certain derogations for big data projects in the context of research and development, scientific and statistical studies. Specific legislation exists in some member states regarding data analytics for public health purposes. However, for other statistic and R&D purposes a number of member states do not provide for any specific exemption.

The question of whether a new purpose is "compatible" with the original purpose is a form of balancing test relying on the "reasonable expectations" of the data subjects and the existence of appropriate safeguards. Recital 50 GDPR provides that:

In order to ascertain whether a purpose of further processing is compatible with the purpose for which the personal data are initially collected, the controller, after having met all the requirements for the lawfulness of the original processing, should take into account, inter alia: any link between those purposes and the purposes of the intended further processing; the context in which the personal data have been collected, in particular the reasonable expectations of data subjects based on their relationship with the controller as to their further use; the nature of the personal data; the consequences of the intended further processing for data subjects; and the existence of appropriate safeguards in both the original and intended further processing operations.

2.7 Data protection impact assessments

Most data lake applications will require a data protection impact assessment under Article 35 GDPR because the processing is likely to result in a "high risk" to the rights and freedoms of individuals. In some cases the data controller(s) may not consider that data analytics creates a "high risk", and that a formal DPIA is not required. Nevertheless, some form of impact assessment is necessary in order to demonstrate that the data controllers implemented appropriate organisational and technical measures taking into account the context and risks of processing (Article 24 GDPR), and that privacy by design principles (Article 25 GDPR) were applied to the data lake.

In theory the security impact assessment will already have been done at the infrastructure stage.

The results of the impact assessment should be communicated to the governance committee, and made available on request to the affiliates that are the original data controllers.

References

1. CJEU Case C-582/14 19 October 2016.
2. Conseil d'Etat, Case n° 393714, 8 February 2017.

Alicante
Amsterdam
Baltimore
Beijing
Brussels
Budapest
Colorado Springs
Denver
Dubai
Dusseldorf
Frankfurt
Hamburg
Hanoi
Ho Chi Minh City
Hong Kong
Houston
Jakarta
Johannesburg
London
Los Angeles
Louisville
Luxembourg
Madrid
Mexico City
Miami
Milan
Minneapolis
Monterrey
Moscow
Munich
New York
Northern Virginia
Paris
Perth
Philadelphia
Rio de Janeiro
Rome
San Francisco
São Paulo
Shanghai
Shanghai FTZ
Silicon Valley
Singapore
Sydney
Tokyo
Ulaanbaatar
Warsaw
Washington, D.C.
Zagreb

Our offices
Associated offices