www.jusletter.ch

Marine Largant

Data Lakes: What are the key issues from a data protection perspective?

Data lakes, these huge repositories of structured and unstructured data are very useful to process data in bulk, conduct analysis at a group level or assess a trend based on a large amount of data. However, as soon as personal data are introduced in a data lake, the data protection requirements must be implemented, which can be quite challenging when faced with an incalculable amount of data and multiple stakeholders and users. We will analyse here some of the data protection aspects which must be taken into account when launching a data lake and during its life cycle.

Category of articles: Articles
Field of Law: Data Protection

Citation: Marine Largant, Data Lakes: What are the key issues from a data protection perspective?, in: Jusletter 23 September 2024

ISSN 1424-7410, jusletter.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

Contents

[1] Data lakes aim at collecting centrally and allow for the processing of a large volume of structured and unstructured data from heterogeneous sources. The data lakes host «big data» which involves large volumes of static or non-static data from various sources, which is stored in an IT infrastructure and processed with software tools to produce insights. Data lakes and big data are the merger of data warehousing, data mining and cloud computing.[1]

[2] We will take as an example in this analysis a group-managed data lake in which group entities upload data collected locally by each of them, based on which the headquarter or certain entities of the group conduct specific analysis for various purposes. The underlying IT architecture can be, for instance, composed of:

- One or several repository data lake(s), on which all data are collected; and
- One or several applications that will connect to the repository data lake(s) and extract the data they need to perform a specific data driven analysis.

[3] In this setup, the data lakes are pure data repositories and the processing is performed by applications (including artificial intelligence systems and software), which will extract data sets for a specific purpose.

[4] Data lakes can be used for instance for the monitoring, control and supervision of the risk at the group level or to analyse consumers' preferences and derive statistics. They are as well used by the banking industry in order to prevent and detect threats and frauds in transactions, or in the area of the scientific research.

[5] Most of the data stored in a data lake are not personal data. However, depending on the purpose of the analysis performed on the data gathered in the data lake, the collection of personal data may be necessary. This is where the data protection regulation enters into play and the data protection principles must be implemented in the data lake and the connected applications.

[6] We will analyse hereafter in which roles, from a data protection perspective, the different stakeholders intervene in the data lake infrastructure (Section 1) and what are the main data protection requirements to implement within the data lake environment (Section 2).

1 Olivier Heuberger, Profiling im Persönlichkeits- und Datenschutzrecht der Schweiz, in: LBR – Luzerner Beiträge zur Rechtswissenschaft Band/Nr. 144, 2020, p. 26–41, p. 26.

1. Definition of the respective stakeholders' roles

1.1. In which capacity is acting each stakeholder?

[7] One of the main challenges of the processing of personal data in data lakes is the definition of the different stakeholders' roles as part of the processing.

[8] Often, different entities of the group and various external service providers will be involved in the process and will take part, at different levels, in the processing. We can distinguish:

- The entity (e.g., headquarter of the group) which decides to create a data lake and as such defines the data lake architecture and functionalities, the type of data it wishes to collect from the local group entities and the purposes of the processing for each application which will extract data from the data lake. This entity acts as the controller of all personal data processed in the global architecture as it defines the underlying purposes and means of the processing in the data lake and applications (the «**data lake controller**»).
- The local entity which collects the personal data originally from the data subjects and transfers it to the data lake. This entity acts as the original controller which has the direct contact with the data subject (the «**original controller**») and not as processor of the data lake controller since generally the collection of personal data from the data subjects is not made on the instruction of the data lake controller but for another purpose specific to the local entity.
- The entity (intra-group or external) in charge of managing the data lake and/or the applications, such as the implementation of the infrastructure, the hosting of the data lake / applications (which can be cloud based), the technical interventions. This entity acts as data processor of the data lake controller as it implements its instructions in the IT infrastructure and provide support on the functioning of the data lake and the applications (the «**data lake processor**»).
- The local entity which will extract personal data from the data lake or use the outcome of the processing by the applications for its own needs. This entity acts as a controller as it will use data from the repository data lake or the analysis made upstream by the data lake controller, in order to fulfil its own purposes (the «**local controller**»).

[9] As such, in a group, a local entity can act both as an original controller (when it transfers to the data lake the personal data collected from the data subjects) and as a local controller (when it uses personal data shared by other group entities to conduct its own analysis).

[10] If the data lake serves the own purposes of both the data lake controller and the local controller and that the latter as well has the capacity to define the purposes and means of the processing in the data lake or the applications, then there could be a joint controller relationship. In such instance, the definition of the data lake's or the applications' architecture or at least of some of their functionalities, comes from a common decisions-making process between the data lake controller and the local controller and each of them assume individually the responsibility for the processing for which it determines the purposes.[2]

[2] Philippe Meier/Nicolas Tschumy, in: Commentaire romand de la Loi sur la protection des données, Helbing 2023, Art. 5 N. 110.

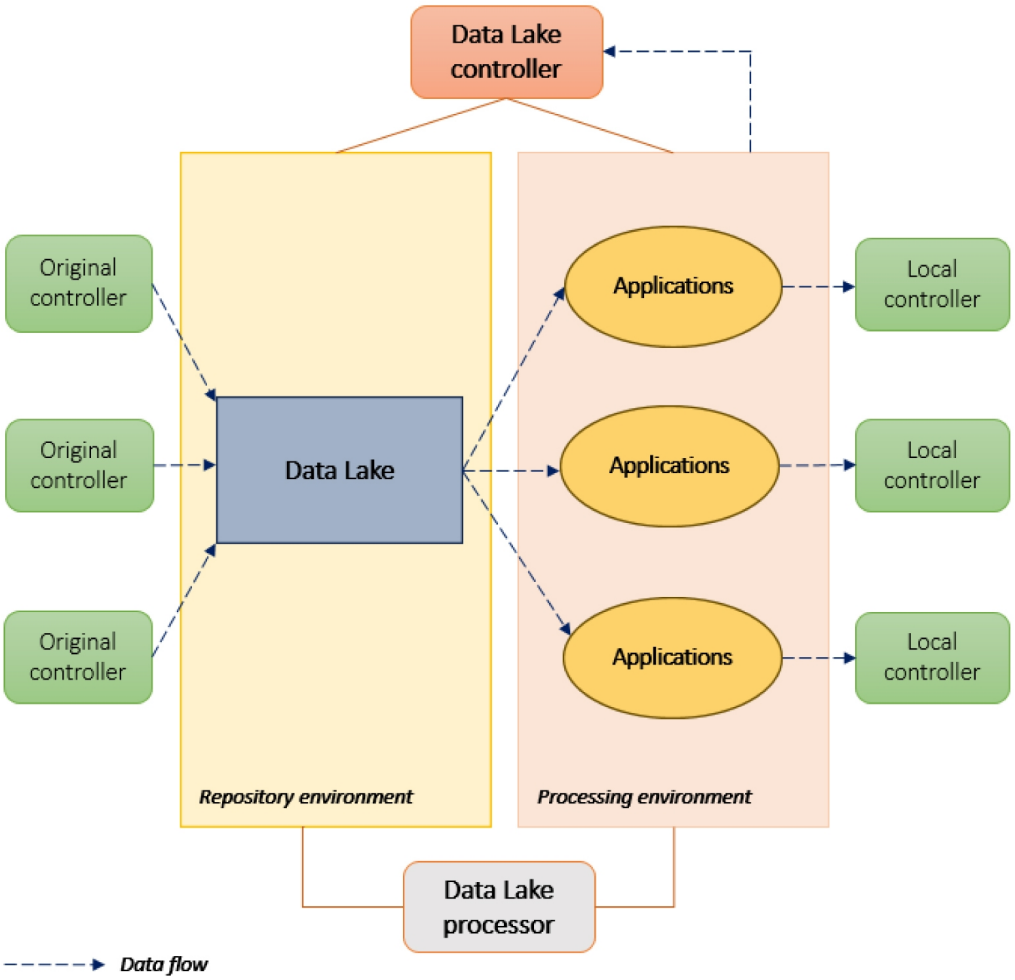
[11] The risk of not defining clearly the role of each stakeholder as part of the data lake infrastructure is to lose the view on which entity controls the data and to mix responsibilities in relation to the personal data processed.

[12] In the context of this analysis, we distinguish mainly two stages in the use of data lakes:

1. Launch phase: when the data lake controller designs and creates the data lake and the connected applications and collects the data from the original controllers in order to run analysis.
2. Development phases: when the data lake controller develops new functionalities in the data lake or the connected applications, or connects new applications to the data lake, or when a local controller requests to initiate a new processing for its own purpose, i.e. all *scenari* which lead to a new collection or processing of data.

[13] The legal and risk assessments, of which we will analyse some components in the developments below, must be performed by the controllers involved at the launch phase, but as well at each development phase which implies a new processing (e.g. new collection of data, new purpose, new transfer of data) that has not been taken into account in the launch phase.

The data lake infrastructure is reflected in the below chart:



1.2. What are the contractual implications?

[14] It is important to define these roles at the outset of participating in a data lake, as the capacity, in which each entity is involved, will determine the data protection contractual framework to put in place with each of them. For instance:

- The data lake controller shall enter into a service agreement and a data processing agreement with the data lake processor (controller to processor relationship);
- The original controller shall enter into a service agreement with the data lake controller (controller to controller relationship);
- The local controller shall enter into a service agreement with the data lake controller (controller to controller relationship).

[15] This documentation can be adopted through a group-wide contractual set-up, which shall then integrate harmonized data protection rules throughout the group (including in relation to the cross-border transfers of personal data, cf. Section 2.3 below), but as well allow for country specific rules, as some original controllers will need to integrate data protection or regulatory requirements to the contractual framework.

[16] Failure to put in place the appropriate contractual framework with the data processor, as required by Article 9 (1) of the Swiss Data Protection Act (DPA)[3] may lead to criminal sanctions of up to CHF 250'000 as per Article 61 (b) DPA.

2. Implementation of the data protection requirements in the data lake environment

[17] The data lake controller must take into account the data protection principles and requirements when building the architecture of the data lake and of the connected applications. It is responsible for the incorporation from the outset of the technical and organisational measures within the IT infrastructure in conformity with the data protection by default and by design principles (Article 7 DPA when the data lake controller is a Swiss entity, respectively Article 25 of the General Data Protection Regulation (GDPR) if the data lake controller is located in the European Union).

[18] We will analyse thereafter the application to the data lake and the connected applications of the main Swiss data protection requirements and their application in particular by an original controller located in Switzerland. Similar requirements are applicable under the GDPR for the group entities located in the European Union.

2.1. Data minimisation and purpose limitation

[19] A main challenge when developing data lakes in groups of entities, where multiple entities are involved as data exporters (e.g. the original controllers transferring personal data to the data

[3] Under the GDPR, this requirement and the content of the agreement with the data processor is provided by Article 28 GDPR.

lake) and data importers (e.g. the data lake controller and the local controllers using the personal data hosted in the data lake), is the clear definition of the purpose of the data lake, respectively of the applications that will perform the processing of data stored in the data lake.

[20] When defining the scope of data to be collected in a data lake (at the launch phase and later on as part of the development phases), the data lake controller must assess what data will be necessary for the foreseen processing in order to avoid collecting a large amount of data that will not be used for the processing. A wide and general definition of the purposes of the processing will open the door to the collection of an extensive amount of data (as such contrary to the principle of data minimisation,[4] as per Article 6 (4) DPA) and to their processing for purposes not properly defined in advance (as such contrary to the principle of purpose limitation,[5] as per Article 6 (3) DPA). If the purposes of the processing are too broadly defined, the data protection and risk analysis performed at the outset of the project will lack specificity and the project will lead to a collection of data for purposes that in reality will be defined in the future. The collection of a large amount of personal data for «to be defined» purposes is specifically what the data protection legislation aims at avoiding. The risk for the data lake controller is notably to assume the liability for the processing of personal data (since the storing, even without further processing, is still considered as processing of personal data) without having a real use of such data. Therefore, the data lake controller should, as much as possible, collect anonymized data and always assess whether the purpose of processing of the applications can be achieved with anonymized data.

Furthermore, the data lake controller should introduce in the launch phase general data protection and security rules in order to regulate the future developments of the data lake and the connected applications, as well as their use by the local controllers (see Section 2.2 below), during the development phases.

[21] Respectively, the original controllers must conduct their own assessment of the legality of the transfer of personal data in the data lake. They shall in this context review both the conformity of the data lake controller's analysis with their local requirements and the legality of the processing of the data they contribute to the data lake for a different purpose than the one they were collected for, as the case maybe.[6] To be able to perform their own assessment of the compatibility of the new purpose with the initial purpose of the processing, the original controllers must enquire from the data lake controller (i) the purpose of the collection of personal data in the data lake and (ii) the purpose of the processing of such data by the applications. If the purposes are too broadly defined by the data lake controller, the original controller cannot perform a precise legal assessment. For a complete assessment, the original controller must obtain and analyse (a) the list of personal data that will be transferred to the data lake, respectively processed by an application, and (b) the purposes of the processing by such application. Such assessment must be performed when an original controller firstly transfers personal data to the data lake (launch phase), but as well each time an application develops a new processing or a new application is connected to the data lake (development phases).

4 As a parallel, the data minimization principle is as well provided by Article 5 (1) (c) GDPR.

5 As a parallel, the purpose limitation principle is as well provided by Article 5 (1) (b) GDPR.

6 In relation to the secondary use of personal data in the context of data lakes, see: Sultan S./Jensen C.D., Secondary Use Prevention in Large-Scale Data Lakes. In: Arai, K. (eds), Intelligent Computing. Lecture Notes in Networks and Systems, 2021, vol 285.

[22] Should the Swiss original controller fail to perform a full legal assessment of the data lake project and its implications on the data subjects and to make sure that the data protection principles provided by Article 6 DPA are fully integrated, it exposes itself to administrative measures from the Swiss Data Protection and Information Commissioner (FDPIC), who may order the modification, suspension or termination of the processing and the deletion of the personal data involved (Article 51 (1) DPA).[7]

[23] Therefore, data lake controllers are faced with a challenge between on the one hand their need for flexibility in using a large data base for processing to be determined in the future and, on another hand, the requirement to specifically define which personal data are necessary for specific purposes.

[24] Technical solutions are developed which allow for the integration in the data lake and in the functionalities of the applications of the purpose limitation principle. Such solutions integrate within data lakes functionalities to allow the discovery and classification of data for a specific purpose, the identification of personal data in the data lake, the matching of data about the same person from different sources or the deletion of personal data from several datasets and as such improve the processing of personal data in conformity with a specific purpose.[8] Such functionalities are integrated through metadata models which enable searching and accessing the data based on specific rules and as such determine and control more specifically the processing made by the applications on the data lake.

[25] This functionality is crucial when the original controller is faced with an access request from a data subject (Article 25 DPA)[9] and must therefore identify notably the personal data of the latter that have been transferred to the data lake and inform her/him on the purpose of the processing. Failure to comply with an access request may lead to criminal sanctions of up to CHF 250'000 as per Article 60 (1) (a) DPA.

2.2. Access limitation

[26] The access limitation principle is part of the technical and organisational measures to be put in place by the controller to ensure the security of the processing (Articles 7 (2) and 8 DPA). It is specifically detailed by Article 3 of the Swiss Data Protection Ordinance (DPO).[10]

---

7 Under the GDPR, the controller has an obligation to document and be able to demonstrate its compliance with the GDPR's principles and requirements. This is the accountability principle provided by Article 5 (2) GDPR. Although such accountability principle is not formally provided in the DPA, the Swiss controller shall still be prepared to demonstrate that it has performed an analysis of the project from a data protection perspective in case of investigation from the FDPIC.

8 Technical solutions, that can be introduced in data lakes, exist in order to link a specific purpose to a data, through the use of meta data providing the corresponding instruction: Rigo Wenning, Big Data und Datenschutz, in: Zeitschrift für Datenrecht und Informationssicherheit, p. 96–99; Olivier Heuberger, Profiling im Persönlichkeits- und Datenschutzrecht der Schweiz, op. cit. N. 46, 47; for an example of a metadata model for the management of metadata see Rebecca Eichler/Corinna Giebler/Christoph Göger/Holger Schwarz/Bernhard Mitschang, Modeling metadata in data lakes – A generic model, in: Data & Knowledge Engineering 136 (2021) 101931.

9 In the GDPR, the right of access is provided by Article 15 GDPR; see as well the trilogy of articles on case law C-154/21 of 12 January 2023 of the European Court of Justice on the access right and its implication in Swiss data protection law: Livio di Tria, Philipp Fischer, 11 April 2023, swissprivacy.law 216, 217, 218.

10 In the GDPR the access limitation is part of the integrity and confidentiality principle provided by Article 5 (1) (f) GDPR, as the controller must put in place appropriate security measures to prevent the unauthorized or unlawful processing of personal data.

[27] However, in the context of data lakes, the limitation of accesses, based on a need-to-know basis, is particularly difficult to implement since a lot of stakeholders are involved in the whole IT infrastructure. The risk is to leave open the accesses to personal data collected in the data lake to all data lake processors and local controllers, independently of their need to access and use such data.

[28] The review and limitation of accesses is therefore relevant for instance when the data lake controller defines to which data from the data lake will an application have access to, or when it receives a request from a local controller to use the data from the data lake for a specific purpose.

[29] Therefore, the data lake controller must put in place a system of access review in order to analyse the necessity and proportionality of accesses to the data lake and the connected applications, as well as a segregation of the data that can be accessed for a specific purpose. As part of this access review system, a local controller which will want to access data from the data lake in order to perform its own processing, or to use outcomes from an application for its own purposes, shall primarily submit a request to the data lake controller. The access request shall identify in particular which personal data the local controller wants to access, for which specific purpose, which processing it will perform and whether the personal data will be transferred to another stakeholder. The data lake controller shall then identify which original controllers are concerned by the request (because personal data they provided are involved) and inform them of the new request. The original controllers concerned shall then perform a new legal assessment (to the extent the access request implies a data transfer that was not analysed as part of the initial legal assessment performed by these original controllers) and approve the processing of personal data by the local controller.

[30] A pyramidal access review system initiated by the local controller and approved by the data lake controller and the original controllers concerned should be implemented in order to ensure that the original controllers are fully aware and remain in control of their personal data flows.

[31] The access granting process and access control system in conformity with Article 3 DPO will be reviewed by the FDPIC in case of investigation and a lack of control or a failure in the system may lead to the administrative measures mentioned under section 2.1 above and to a criminal sanction of up to CHF 250'000 as per Article 61 (c) DPA.

Excursus

[32] A data lake can be used at a group level in order to issue consolidated regulatory, statutory or risk reports in relation to the activity of the group entities. Such processing is performed by an application which will use data from the data lake to perform the analysis and generate the reports. When the reports created are mandatory for the group, they will be transferred to foreign authorities. This can be for instance common as part of international banking groups.

[33] Independently from the presence of personal data within such report, the transfer of Swiss entities' information to foreign authorities may trigger issues in relation to the Swiss blocking statutes, provided by Article 271 of the Swiss Criminal Code (SCC). This Article prohibits any person to carry out or facilitate activities on behalf of or for a foreign State on Swiss territory without lawful authority. This disposition is designed to prevent a foreign State from exercising its authority on Swiss territory, and thus protects Swiss sovereignty. Procedures designed to circumvent the channels of criminal and administrative mutual assistance typically fall within the scope of this provision. Consequently, the transfer from Switzerland of information and doc-

uments which, under Swiss law, can only be transmitted abroad by order of a Swiss authority, infringes the legal asset protected by Article 271 (1) SCC. The case law has specified that only information that is freely at the disposition of the person providing it may be transferred without prior authorization. This is not the case for data concerning third parties, such as the company's customers.[11]

[34] In such cases, the Swiss original controller must carefully analyze before transferring data in the data lake the potential application of Article 271 SCC[12], which may apply even if the information is provided through the intermediary of another entity of the Group, on top of the implications in terms of data protection if personal data are involved.

2.3. Cross-border data transfers

[35] In conjunction with the management of accesses to the data lake and the applications is the analysis of the cross-border data transfers (i) to countries that do not provide for an adequate level of data protection from a Swiss law perspective (i.e. outside of the list of countries provided by the Annex 1 to the DPO) and (ii) to countries outside of the original controller's jurisdiction in case of secrecy limitation applicable to the personal data (professional or banking secrecy if applicable). This analysis must be performed at the group level by the data lake controller during the launch phase of the data lake, as well as during the development phases when an application performs a new processing which may imply a cross-border data transfer. Since most of the data processing is performed intra-group in the context of the data lake, the data lake controller may want to implement group-wide solutions for the transfer of personal data between the group entities (e.g. binding corporate rules, harmonized process and contractual framework for the cross-border transfer of personal data). The data lake controller should as well implement the required contractual framework and the appropriate technical and organizational measures when personal data are accessed outside of the group, for instance by a third party data lake processor.

[36] The original controller must as well conduct its own assessment of the cross-border data transfer and carefully review the legality of the transfer of personal data to the data lake controller, but as well to any other stakeholder that will have access to the personal data later on. Indeed, the original controller is responsible for the compliance with the requirements for the communication of personal data abroad and in particular for the implementation of the appropriate contractual guarantees and security safeguard in this context (Articles 16 and 17 DPA).[13] The original controller must therefore analyse as part of its legal and risk assessment the whole data lake's and applications' infrastructure and identify the accesses involved, the contractual documentation in place and the technical and organisational measures implemented in order to assess their conformity with the Swiss data protection and regulatory (as applicable) legislation.

[37] Furthermore, a failure to comply with the requirements of Articles 16 and 17 DPA is subject to criminal sanctions of up to CHF 250'000 as per Article 61 (a) DPA.

---

11 Swiss Federal Court decision of 1 November 2021, 6B_216/2020, commented by K. Villard, Transmission de données clients aux États-Unis : Condamnation d'un gérant de fortune, in: CDBF 1211, 26 November 2021.

12 See Philipp Fischer/Deborah Hondius, Article 271 (1) of the Swiss Criminal Code: myth or reality?, in: Jusletter vom 4. April 2022, regarding the constitutive elements of Article 271 (1) SCC.

13 Under the GDPR, the requirements in relation to the transfer of personal data in the absence of adequacy decision from the European Commission are provided by Articles 46 to 49 GDPR.

2.4. Retention periods

[38] Another challenge presented by the storing of personal data in a data lake is the retention period of the personal data hosted in the repository. Whereas the data lake controller will want to keep the data for as long as possible in order to perform further analysis, the original controllers will want to apply their own data retention periods, specific to their applicable regulations and to the data sets they transfer to the data lake. The data lake controller must therefore be able to apply specific retention periods to segregated data sets in order to comply with the local requirements of the original controllers, or setting a reduced retention period for the data provided by all group entities.

[39] In the same way, the original controller shall remain in control of the personal data it transfers to the data lake and as such be able to cease the transfer, respectively request for the anonymisation, retrieval or deletion of the personal data already transferred[14] (in order to comply with Article 6 (4) DPA).[15] In the same way, the original controller must be able to implement a request of erasure from a data subject and if requested by the latter ask the data lake controller to delete all data concerning the individual in the data lake (Article 32 (2) (c) and (4) DPA).[16] Failure to comply with these requirements may lead to the administrative measures mentioned under section 2.1 above.

2.5. Information of data subjects

[40] If the purpose of processing by an application is not linked to the purpose for which the original controller collected the personal data, it shall inform the data subjects of such new purpose in compliance with the duty to inform of the controller as per Article 19 DPA,[17] for instance by updating the data protection policy.

[41] For instance, a data lake can be used for the management of the human resources data base at the group level. In this context, the employees' data protection notice shall provide for the transfer of employees' personal data to other group entities for such purpose.

[42] The original controller is primarily responsible for the duty to inform the data subjects, as it has the direct contact with them. Therefore, the data lake controller and the local controller should delegate their information duty towards the data subjects to the original controller.

[43] This is particularly important for the Swiss original controller to comply with its duty to inform the data subjects as this requirement is sanctioned by a fine of up to CHF 250'000 as per Article 60 (1) DPA.

---

14 For an example of a metadata model allowing the personal data discovery, anonymization or deletion, see: D. Oreščanin/T. Hlupić/B. Vrdoljak, Managing Personal Identifiable Information in Data Lakes, in: *IEEE Access*, 2024, vol. 12, pp. 32164–32180.

15 Under the GDPR, the storage limitation principle is provided by Article 5 (1) (e) GDPR.

16 Our understanding is that the communication by the controller of the deletion request to the recipients of the personal data is not automatic under the DPA, as it is under the GDPR (as per Article 19 GDPR in relation to Article 17 GDPR), but can be specifically requested by the data subject under Article 32 (4) DPA.

17 The information duty is provided by Articles 13 and 14 GDPR.

3. Conclusion

[44] Although the introduction of the data protection requirements is a technical challenge for the data lake controllers, which additionally may enter in conflict with their interest to the processing, it is crucial to take them into account from the outset of the creation of the data lake, respectively as part of its evolution and improvements, including those of the applications which process the data from the data lake.

[45] The original controllers shall be closely involved into the legal and risk assessment of the data lake infrastructure as they bear the original responsibility of the compliance with data protection rules towards the data subjects.

Me Marine Largant is an associate at OBERSON ABELS SA. After starting her career with a Big4, she is now a member of the Geneva Bar, specializing in corporate law and data protection.