Mapping Identity Management in Data Lakes

Jan Zibuschka[1], Lothar Fritsch[2]

Abstract: Data lakes are an emerging paradigm for large-scale, integrated data processing within organizations. While it has been noted in earlier work that data governance is central for the successful operation of a data lake, and that privacy is a central issue in such a setting as personal information may be processed, the governance of personal information in data lakes has received only cursory attention. We propose tackling this information using identity management functions and perform a systematic gap analysis based on the FIDIS typology of identity management systems.

Keywords: identity management; data lake; privacy; data governance; data protection

1 Introduction

Modern organizations experience an influx of digital information from various sources, which they analyze as part of a multitude of business processes, based on the massive computing resources available in contemporary Clouds. These data analytics have become a mainstay of value creation in the information economy [Ma16]. data lakes are quickly becoming the dominant paradigm for comprehensive integration of this data. While there are various implementations and architectures, covering the different relevant business processes to varying degrees [ML16, Na19], in general data lakes facilitate the integration of information from distributed, heterogenous sources within an organization, and allow for performing a wide array of analytics on the information [Gi19]. As the information in data lakes is drown from diverse, heterogenous sources, in various formats, and may or may not be accompanied by useful annotations [Na19], in its initial state it is more of a "data swamp" [Gi19], with disparate, unconnected information and formats, and may remain so if there is no prudent data governance [GH19].

One important societal challenge of Big Data analytics in general are the privacy issues they induce. Data analytics unleashed on a broad basis of information can lead to insights about a data subject that the data subject may not have foreseen, which will lead to objections down the line [Ma16]. While for dedicated processing of big data in a fully controlled, homogenous environment, various approaches for privacy-preserving analytics exist [Me16], these do not translate to the more decentralized processing

[1] Robert Bosch GmbH, Zentralbereich Forschung und Vorausentwicklung, Renningen, 70465 Stuttgart, jan.zibuschka@de.bosch.com

[2] Oslo Metropolitan University, Pilestredet 35, 0166 Oslo, lothar.fritsch@oslomet.no

embodied by heterogenous data storage and processing in data lakes. Therefore, privacy engineering for managing personal information on a data lake level would once again focus on data governance of personal information, providing an infrastructure promoting transparency and accountability [Sp19], which are protection goals of the European General Data Protection Regulation (GDPR) [Sp19], and are also commonly used for privacy in data analytics on other continents [We07].

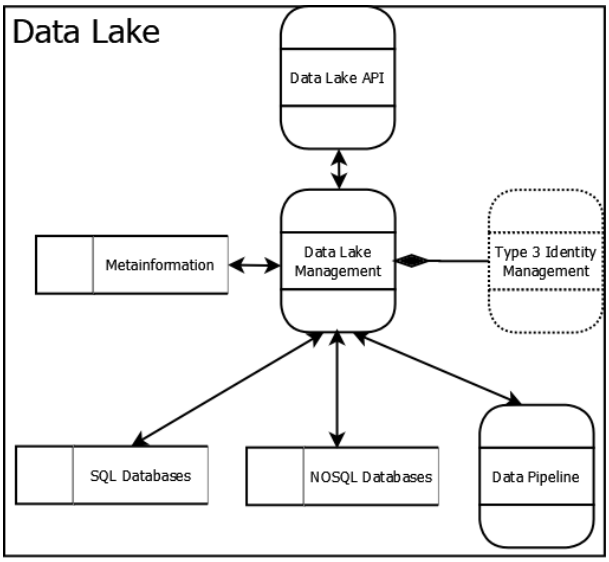

Fig. 1: Structural components of a data lake, proposed identity management

To put it in data lake terminology: we cannot have a data swamp of personal information in an organization but need a more structured approach. The structured processing and transmission of personal information is in the domain of identity management (IdM) systems. While some identity management functions for administration are present in commercial data lakes [Kl17], and identity management has also been acknowledged as a core function of data lake management systems in research [Na19], there has not, been a systematic investigation of the topic. This contribution aims to fill this gap. Based on an analysis of the flow of personal information in a generalized data lake architecture, we identify key identity management functions in the processing of personal information in a data lake based on the FIDIS (Future of Identity in the Information Society, and EU-funded project[3]) typology of identity management systems [Ba05].

[3] http://www.fidis.net/

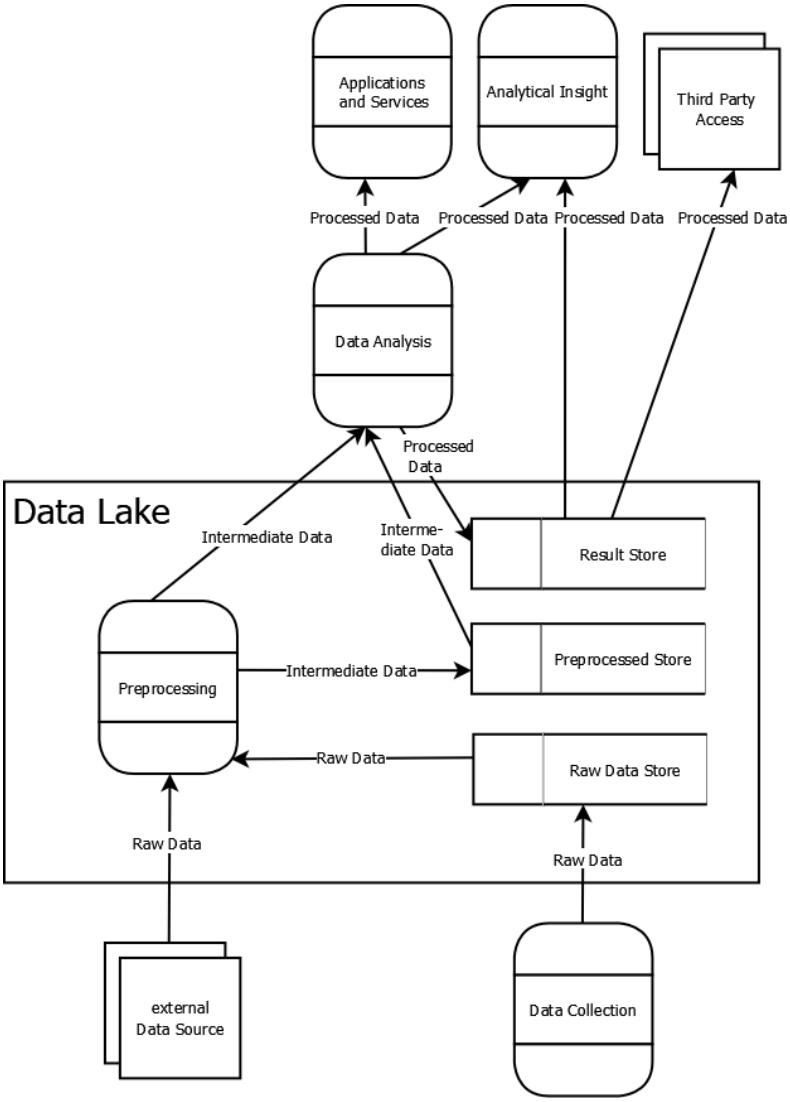

Fig. 2: Data lake reference architecture with personal information flow

2 Reference Architecture for Data Lakes

As a basis for mapping identity management in data lakes, we build a simplified reference architecture based the data lake architectures and processes described in related work. Our aim for this architecture for it to be universally applicable and comprehensive with regards to the covered business processes. Firstly, we establish a top-level, structural view of what constitutes a data lake. data lakes comprise various data stores [Na19], including SQL databases, NOSQL databases, and data processing pipelines [GH19] (See Fig. 1). Those data stores are orchestrated by one or several data lake management components, which expose a control interface via an API [Be17]. This data lake management draws on a base of metainformation [GH19] about the underlying data sources. In addition, we map the flow of personal information though a data lake, see Fig. 2. The main structural steps of personal information processing in a data lake are:

1. Personal information may enter such a data lake from external data sources, or from data collection performed by the organization operating the data lake [GH19]. A combination of both approaches is also possible, as externally acquired information may require annotation [GH19]. In any case, the raw data acquired is usually stored in the data lake [Gi19] in an untransformed state [KW18].
2. The information may then be preprocessed in the data lake. Operations that are performed on this level include data extraction and data cleaning [Na19], with the overall aim of bringing the information in a harmonized state fit for further processing. The results of this preprocessing may again be stored in the data lake [Gi19], or may be generated on the fly.
3. Analysis of the information is performed. This step may be performed within the data lake, i.e., in the case of integrated processing pipelines [GH19], or may take place outside of the oversight of data lake management [ML16]. We denote this with a dotted, extended border for the data lake system. Whether the processing is done in the data lake or not, results may be stored there for archival purposes [Gi19].
4. Finally, access to the result of processing may be given to data analysts or services internal to the organization operating the data lake, or to external entities [ML16].

Thus, analysis of the flow of personal information in data lakes can differentiate input, preprocessing, analysis, and output stages of processing. Fig. 2 gives an overview of our data lake reference architecture for analyzing personal information flow.

3 Types of Identity Management in Data Lakes

3.1 FIDIS Identity Management Typology

In addition to those processing stages, we build on the FIDIS classification of identity management systems [Ba05]. The typology differentiates type 1 identity management systems for account management, that implement authentication, authorization, and access control; type 2 identity management systems for processing of user data by an organization, and type 3 identity management systems that enact user control of their identity information and pseudonyms that are exposed.

A data lake processing personal information would be considered a type 2 identity management system as a whole under this definition. Thus, our contribution is concerned with tracing relevant junctions for the introduction of type 1 and 3 identity management in this overall type 2 identity management structure, and other type 2 identity management structures within the organization that may connect to the data lake.

Overall, type 3 identity management is the main gap identified in our analysis, which is significant, as type 3 identity management most directly addresses legal requirements [Sp19]. Type 1 identity management offers an underlying infrastructure allowing only authorized access, which is necessary for selective transparency and intervenability, and to implement state of the art security. An overview of the position of these components is provided in Fig. 3. The following sections provide some more in-depth discussion of the types of identity management in data lakes.

3.2 Type 1 Identity Management

With regards to type 1 identity management components, it is notable that the subjects of the identity management systems can vary quite significantly. Type 1 identity management in data lakes may concern the end user whose personal information is processed, system administrators, data scientists accessing the data lake, or even the organization operating the data lake itself.

This is for example relevant in the integration of external data sources. An external data source may require authentication to access. Then, it be integrated in the data lake using an organizational account representing a link between the organization operating the data lake and the organization operating the external data source on a B2B (business to business) level. Examples include external services offering specialized global information such as weather data, stock market information, or geographic information. The external data sources may, however, also be personal, linked to data lake using an end user account. This may be relevant for individual information such as social network accounts or individual devices and might be accomplished using protocols such as OAuth.

It is notable that type 1 identity management also potentially exists in many of the individual data stores within the data lake, but may not support the authentication needed for processing in the data lake, either being to rigid, not allowing for integrated processing of the data, or not having access to information about the broader processing, such as the accessing entity, that would be needed for a sensible authentication.

This distinction between type 1 identity management for various stakeholder groups is highly relevant in controlling access to the data stored in the data lake. Administrators logging into the underlying systems where the information is stored may get access to it. Analysts of the organization operating the data lake need to be authenticated when accessing the stored information. However, we note the currently documented data lake architectures and management services do not implement any type 1 identity management for the data subject, and therefore cannot offer self-service access, even though it is encouraged by the GDPR [Sp19], specifically regarding access to stored personal information. All in all, data lakes need and, in many cases, already offer on their management layer type 1 identity management for all involved stakeholders, as indicated in Fig. 3.

3.3 Type 2 Identity Management

As already mentioned, the processing of personal information in a data lake as a whole constitutes type 2 identity management in the sense of the FIDIS typology. It is notable that the type 2 identity management in the organization stretched beyond the data lake proper, both into data collection and into further data processing for analytics or services. These steps are also part of our reference architecture for this reason. However, this contribution is concerned with analyzing identity management within a data lake. Data Analytics may be an integral part of a data lake in cases such as integrated data processing pipelines for high throughput Big Data [GH19], bringing this case of type 2 identity management into the scope of our analysis, as depicted in Fig. 3.

3.4 Type 3 Identity Management

While type 3 identity management is part of contemporary data lakes in the form of integration of external data sources that may be under the control of the user, or solely generated by the user [GH19], self-service components in these data lakes do not address the data subject, and instead focus on internal stakeholders, such as data analysts or system administrators [GH19].

While some of the functions associated with contemporary data lake management can constitute type 3 identity management if leveraged by a data subject, such as the ability to identify personal information in a data lake [GH19], the lack of underlying type 1 functions for data subjects makes it unlikely that those functions would ever be used in that way, except when stakeholder roles overlap, e.g., a data analyst is also a data subject.

4 Proposed Extension of Identity Management in Data Lakes

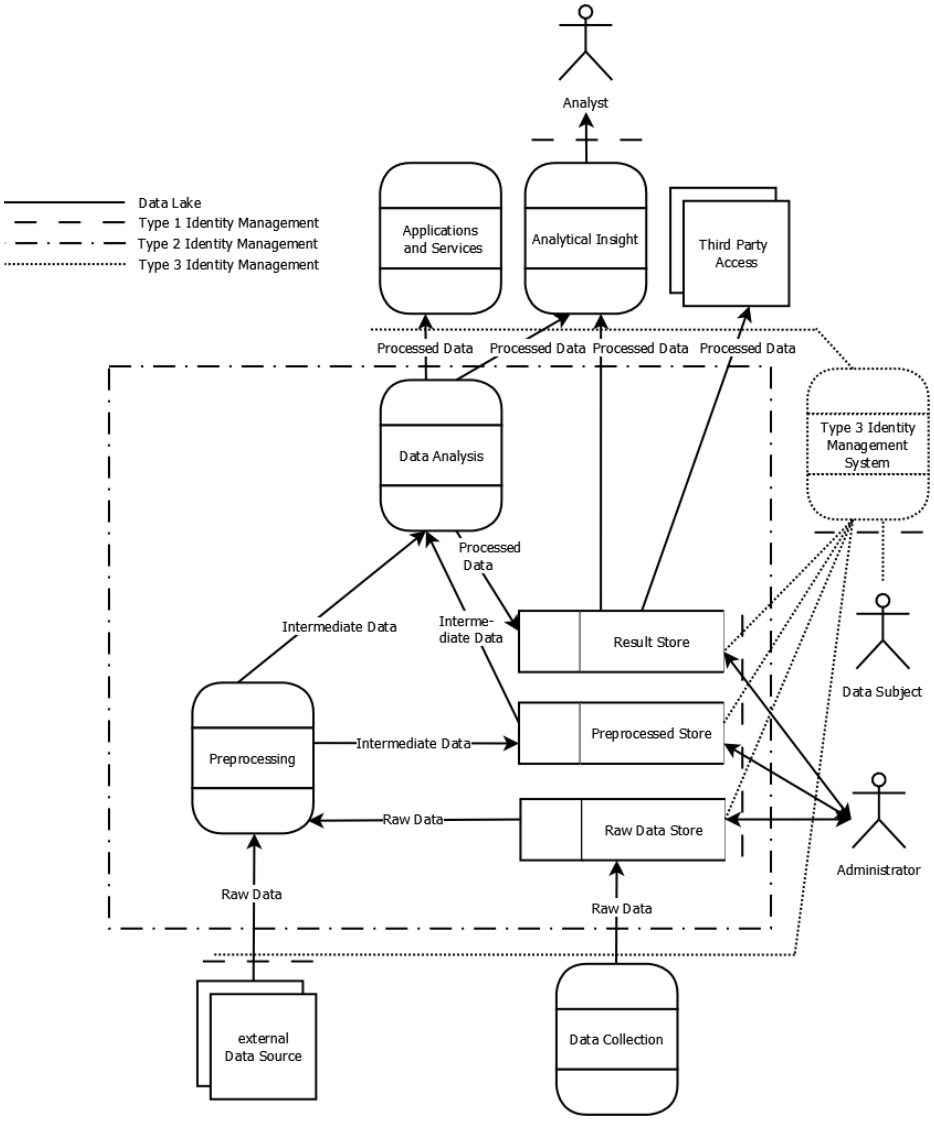

Fig. 3: Identity management in data lakes including proposed type 3 extension

We identify introducing a type 3 identity management component into data lakes, and specifically into data lake management services, as a clear gap of the current implementation landscape. The data lake is linked to data collection, data storage, large-scale data processing, and access to results of such processing by value-added services, analysts, and third parties, and as of such offers excellent opportunities to enact intervenability, transparency, and accountability. Further, this would require type 1 identity management for the data subject linked to the type 3 identity management process, to enable self-service transparency and intervenability.

We propose also linking such a component to the data collection processes, allowing for control of the data flowing into the data lake by the Data subject. Finally, we also propose linking the type 3 identity management to the outbound data flow connecting the data lake to applications, services, and analytics access by data analysts. Such a setup would allow for capturing the data subject's consent for which data is collected in the data lake, and for which purposes this data is used. How to store this consent and interlink inbound and outbound data flows we leave as future work.

5 Discussion

In this short contribution, we could only give a very coarse overview of the challenges of privacy and identity management in data lakes. From earlier work, we know that consent management is a very important aspect of implementing privacy in scenarios with multilateral processing of personal information [Ra07, Zi07]. This is even more central in the context of modern data lakes, as the implicit assumption is that data from heterogenous data sources originating from equally heterogenous processing purposes may be integrated. However, this aspect goes beyond the FIDIS typology we build on for this work.

Similarly, we cannot cover the details of authentication to the system. There are several directions for future work in this area. The data subject might be authenticated using federated identity management [Hü10], allowing for cross-domain single sign on and potentially linking to the origin of data in the data lake. That same cross-organizational link may also be established with federated identity management on a B2B level [Hü11]. The data analyst, on the other hand, may require stronger authentication, with might be provided by interoperability of the data lake with a strong authentication infrastructure such as enterprise smart cards [RZ06]. The data analyst may also be given a restricted view of the data, and further privacy protections may be implemented in the system [Gu17]. For example, earlier work has proposed automatically identifying personal information in the data lake [GH19].

It is notable that, from a privacy perspective, interlinking personal information in data lakes is not all bad. The data subject may even benefit from the integration of personal information in a data lake, as raw data, preprocessed information and results of processing, can all be accessed to offer transparency on the level of detail that is most

opportune and understandable. The data subject can also, using identity management functions, potentially control information flowing into the data lake. This includes personal information from both data collection processes of the operating organization and from external data sources. It can also control the use of this information in analytics and services, both internally and at third parties.

All those directions for future work notwithstanding, analyzing identity management in data lakes using the FIDIS typology can make a significant contribution to data governance, as we illustrated in this paper. The reference architecture we also provided may prove useful in the pursuit of any of those possible directions for follow-up. Data lakes are quickly moving from an academic idea to a practical reality. This first contribution could not fully explore the depths of the privacy and identity management challenges that go along with that development. We do, however, encourage future work, which we hope this contribution will inform and motivate.

Bibliography

[Ba05] Bauer, M. et al.: FIDIS Deliverable D3. 1–Structured Overview on Prototypes and Concepts of Identity Management Systems, http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp3-del3.1.overview_on_IMS.final.pdf, (2005).

[Be17] Beheshti, A. et al.: CoreDB: a data lake Service. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2451–2454 Association for Computing Machinery, New York, NY, USA (2017).

[Gi19] Giebler, C. et al.: Leveraging the data lake: Current State and Challenges. In: Ordonez, C. et al. (eds.) Big Data Analytics and Knowledge Discovery. pp. 179–188 Springer International Publishing, Cham (2019).

[GH19] Gröger, C., Hoos, E.: Ganzheitliches Metadatenmanagement im Data Lake: Anforderungen, IT-Werkzeuge und Herausforderungen in der Praxis. In: BTW 2019. Gesellschaft für Informatik, Bonn (2019).

[Gu17] Gursoy, M.E. et al.: Privacy-Preserving Learning Analytics: Challenges and Techniques. IEEE Transactions on Learning Technologies. 10, 1, 68–81 (2017).

[Hü10] Hühnlein, D. et al.: Diffusion of Federated Identity Management. Sicherheit 2010. Sicherheit, Schutz und Zuverlässigkeit. (2010).

[Hü11] Hühnlein, D. et al.: Skidentity – Vertrauenswürdige Identitäten für die Cloud. DA-CH Security. 296–304 (2011).

[KW18] Khine, P.P., Wang, Z.S.: Data lake: a new ideology in big data era. In: ITM web of conferences. p. 03025 EDP Sciences (2018).

[Kl17] Klein, S.: Azure data lake Store. In: IoT Solutions in Microsoft's Azure IoT Suite. pp. 143–154 Springer (2017).

[ML16] Madera, C., Laurent, A.: The next information architecture evolution: the data lake

wave. In: Proceedings of the 8th International Conference on Management of Digital EcoSystems. pp. 174–180 (2016).

[Ma16] Mai, J.-E.: Big data privacy: The datafication of personal information. The Information Society. 32, 3, 192–199 (2016).

[Me16] Mehmood, A. et al.: Protection of big data privacy. IEEE access. 4, 1821–1834 (2016).

[Na19] Nargesian, F. et al.: Data lake management: challenges and opportunities. Proc. VLDB Endow. 12, 12, 1986–1989 (2019).

[Ra07] Radmacher, M. et al.: Privatsphärenfreundliche topozentrische Dienste unter Berücksichtigung rechtlicher, technischer und wirtschaftlicher Restriktionen. 8. Internationale Tagung Wirtschaftsinformatik 2007 - Band 1. 237–254 (2007).

[RZ06] Roßnagel, H., Zibuschka, J.: Single-sign-on mit Signaturen. Datenschutz und Datensicherheit-DuD. 30, 12, 773–777 (2006).

[Sp19] Spagnuelo, D. et al.: Accomplishing Transparency within the General Data Protection Regulation. In: Proceedings of the 5th International Conference on Information Systems Security and Privacy. pp. 114–125 (2019).

[We07] Weitzner, D.J.: Beyond Secrecy: New Privacy Protection Strategies for Open Information Spaces. IEEE Internet Computing. 11, 5, 96–95 (2007).

[Zi07] Zibuschka, J. et al.: Privacy-friendly LBS: a prototype-supported case study. AMCIS 2007 Proceedings. Paper 40. (2007).