WHO OWNS DATA IN THE ENTERPRISE? RETHINKING DATA OWNERSHIP IN TIMES OF BIG DATA AND ANALYTICS

Research paper

Fadler, Martin, University of Lausanne, Lausanne, Switzerland, martin.fadler@unil.ch

Legner, Christine, University of Lausanne, Lausanne, Switzerland, christine.legner@unil.ch

Abstract

Today, a myriad of data is generated via connected devices and digital applications. With recent advances in artificial intelligence (AI), companies are seeking new opportunities to monetize data. This goes along with improving their capabilities to manage big data and analytics (BDA). A critical factor that is often cited concerning the 'soft' aspects of BDA is data ownership, i.e. clarifying the fundamental rights and responsibilities for data. Scholars have investigated data ownership for operational systems and data warehouses, where the purpose of data processing is known. In the BDA context, defining accountabilities for data ownership is more challenging, because data is stored in data lakes and used for new, previously unknown purposes. Based on insights from three case studies with extensive experience in BDA, we identify ownership principles and three data ownership types: *data*, *data platform*, and *data product*. By redefining the concept of data ownership, our research answers fundamental questions about how data management changes with BDA, extending existing concepts on data ownership and contributing to the data governance literature.

Keywords: Data ownership, Data governance, Big data and analytics, Data lake, BDA

1 Introduction

There is no doubt that data is leading to a rising new economy (The Economist 2017) and is fundamentally changing how business is conducted (Davenport et al. 2012; Wamba et al. 2015). With decreasing computing costs and the myriad of data generated via connected devices and digital applications, enterprises are seeking opportunities to improve existing processes and products as well as to develop new data-driven business models (Wixom and Ross 2017). This goes along with improving their capabilities to manage big data and analytics (BDA). A cornerstone of BDA is data lakes, which store large volumes of data in various formats and enable innovation through data exploration and experimentation (Farid et al. 2016; Madera and Laurent 2016; Watson 2017). The business potential of data scales through its inherent characteristic of being nonrivalrous. In contrast to other economic goods, data can be used for multiple purposes at the same time. This idiosyncrasy leads to complexity in data ownership. Data ownership clarifies fundamental rights and responsibilities for data (Hart 2002) and is commonly considered to be beneficial. Grover et al. (2018) emphasized: *"[…] governance that delineates responsibility and accountability for data, [is a catalyst] for BDA value creation"* (p. 417). However, the related debates in practice and research view the concept from different, often contrasting perspectives. More recently, data ownership is seeing increasing interest in public debates. For instance, some governments are introducing privacy regulations to give individuals more rights and to control businesses' uses of personal data (Labadie and Legner 2019). Economists are investigating how data ownership affects social welfare (Jones and Tonetti 2019). In the enterprise context, data ownership is often cited as a critical factor concerning the 'soft' aspects in the creation and use of enterprise data, specifically BDA. As data ownership clarifies fundamental rights and responsi-

bilities, it also underpins data governance (Loshin 2001; Winter and Meyer 2001). Data ownership has been discussed since electronic data processing began (Maxwell 1989; Spirig 1987; Wang et al. 1995). The focus of the subsequent debates has been on data ownership for operational systems and data warehouses, where the purpose of data processing is known. Different scholars have emphasized that data ownership remains important to gain business value from big data (Alexander and Lyytinen 2017; Comuzzi and Patel 2016; Grover et al. 2018). While we can assume that assigning accountabilities for data is still beneficial in today's corporate environment, practitioners emphasize that data lakes require a different approach to data governance (Chessell et al. 2018). These developments raise the question how we need to reinterpret and apply data ownership concepts so as to cope with emerging challenges in BDA environments.

To address this gap, our objective is to understand how data ownership concepts change in the context of BDA. Thus, we ask:

RQ: *How do enterprises define and adapt data ownership in the big data and analytics context?*

We opt for an explorative research design based on multiple case studies (Benbasat et al. 1987; Yin 2003). From the analysis of three companies with significant BDA experience, we identify data osnership principles and three data ownership types: *data*, *data platform*, and *data product*. Our findings extend the existing data ownership concept by integrating the *data platform* perspective, which serves as the required mediator between data supply (*data*) and data demand (*data product*) in BDA environments. Our insights into ownership contribute to the data governance literature generally, particularly to structural aspects of data governance according to Tallon et al. (2013). Based on Grover et al.'s (2018) research framework, they lay the foundation for *BDA governance to facilitate the value creation process*.

The remainder of this paper is structured as follows: We start by reviewing the research field on data ownership and outline the research gap. We then motivate our qualitative research approach and provide an overview of the research process. Third, we present each case in detail. Based on our cross-case analysis, we synthesize our findings into four propositions. We conclude with a summary and discussions of our contributions as well as an outlook on future research.

2 Background

Data ownership is grounded in the general concept of ownership, which is a fundamental mechanism in our society and can relate to different theoretical lenses. There has been research into data ownership since the early days of electronic data processing, and different paradigms can be applied to determine who could or would be entitled to claim ownership of data. In the enterprise context, data ownership principles have been studied for operational systems and data warehouses. With BDA, data is used for new, previously unknown purposes and is stored in data lakes so as to enable data exploration and experimentation. This required that we revisit the data ownership concept.

2.1 The concept of ownership

Ownership is a fundamental concept that is grounded in our everyday life and in fundamental mechanisms of society (Shleifer 1998). It denotes the assignment of rights and responsibilities for a property to an individual or an organization. Three types of rights can be distinguished (Hart 2002): the rights to use, to control, and to remain in control. Concerning the philosophical assumptions, different theories allow one to explain the emergence and assignment of ownership (Hart 2002):

- In the view of *first occupancy theory* (Immanuel Kant), ownership is assigned to the first who possesses a property.
- According to *labor theory* (John Locke), ownership is assigned according to the extent of value added through labor.

- In *utility theory* (Jeremy Bentham and John Stuart Mill), ownership is assigned in the way that it maximizes the benefits for all persons involved.
- In *libertarian theory* (Robert Nozick and John Rawls), ownership must be allocated in ways that do not limit the freedom of others to act autonomously.
- *Personality theory* (Georg Wilhelm Friedrich Hegel) determines ownership by a person's will to invest in an object, which makes them this object's owner.

These different perceptions of ownership also exist concerning data, but must consider data's inherent characteristics, such as nonrivalrousness.

2.2 Data ownership and its paradigms

Generally, data ownership is a *"[…] control issue – control of the flow of [data], the cost of [data], and the value of [data]"* (Loshin 2001, p. 28). In their seminal paper, Van Alstyne et al. (1995) distinguished between data usage rights and ownership of data: *Usage rights* denote the ability to access, create, standardize, and modify data as well as all intervening privileges; *Ownership* implies that the residual right of control is the right to determine these privileges for others.

While owner rights are easily assigned, it is hard to link responsibilities to roles, owing to data's inherent nonrivalrousness (Jones and Tonetti 2019). In contrast to other economic goods, data can simultaneously be used for multiple purposes. This idiosyncrasy has led to debates about data ownership and the uses and control of data (Hart 2002). Generally, responsibilities can be diverse depending on its context of use. Thus, Loshin (2001) explored different data ownership paradigms. Although Loshin (2001) followed a fairly pragmatic approach, the suggested paradigms can be linked to the discussed ownership approaches from philosophy and can help us to understand the complexity as well as to structure the research field (see Table 1). We classify the paradigms according to the socio-organizational context into three categories: individual, organizational, and shared ownership (everyone). We will now present each category.

Data ownership is increasingly being claimed by individuals as the subjects of data (*subject as owner*). With the Internet, personal data is being collected, used, and even sold in nontransparent ways. Thus, the private ownership paradigm often emerges as a reaction once the data collection has been unveiled. This was the case in the Cambridge Analytica scandal, where data of millions of Facebook users was used without their official consent (Confessore 2018). Governments react to these developments by enforcing *individual data ownership* rights with data protection policies such as the General Data Protection Regulation (GDPR) in the European Union. With the emergence of the Internet of Things (IoT), the debate about individual data ownership has gained a new facet, because it remains unclear who owns personal data produced by machines (Janeček 2018). For instance, the data collected by smart meters enable electricity providers to optimize their network and service offerings, but also unveil highly sensitive data about private households, which can easily be misused (McKenna et al. 2012).

In the context of organizations (*enterprise as owner*), the data ownership concept is getting more complex as a result of distributed data creation and processing in organizations (Van Alstyne et al. 1995). Here, three reasons for claiming ownership can be distinguished. First, organizations claim ownership owing to monetary factors of funding (*funding organization as owner*) or purchasing/licensing data (*purchaser/licensor as owner*). These paradigms always involve two parties. On the one side, the organization that funds the party who creates data; on the other side, the organization that purchases or licenses data owned by another party. While in the first case data ownership is transferred to the funding organization without any restrictions, in the second case, data ownership is transferred to the purchasing/licensing party under certain restrictions. Second, an organization may claim ownership by using data. This is typically the case for consuming parties (*consumer as owner*) that require high confidence in the data and therefore take over accountability. It may also apply to parties who read data

from different sources (*reader as owner*) to create or add these to their knowledge base. Third, organizations create business value through data processing and therefore claim ownership. Four paradigms can be distinguished depending on the processing type: creating data (*creator as owner*) or formatting data (*packager as owner*) for a certain purpose, compiling information from various data sources (*compiler as owner*), and decoding data (*decoder as owner*).

The socio-organizational context	The data ownership paradigm (Loshin 2001)	Example	The related philosophical perspective on ownership (Hart 2002)
Individual	Subject as owner	A private person accuses a company of selling his or her personal data to a third party	Libertarian theory: Ownership does not limit the freedom of others
Organization	Creator/Generator as owner	A research firm invests in collecting qualitative data for a market study	First occupancy theory: Ownership by being the first to possess an object
	Consumer as owner	A sales team uses customer phone numbers that are essential for its daily operation	
	Reader as owner	A consultancy collects information on industry trends to extend its knowledge base	
	Enterprise as owner	An enterprise creates, processes (adds value), and distributes data about its products	Labor theory: Ownership through value adding, either by own labor or owning labor Personality theory: Ownership through personal will
	Funding organization as owner	A company pays a research company to collect panel data	
	Purchaser/Licensor as owner	A company buys an address list of potential customers	
	Compiler as owner	A business intelligence department builds a central data warehouse	
	Packager as owner	A web agency designs and formats a web page for a customer	
	Decoder as owner	A company synthesizes information from DNA data	
Everyone	Everyone as owner	A crowdsourced collection of geo-information in a public database	Utility theory: Ownership maximizes the benefits for all involved parties

Table 1 Data ownership paradigms and discourses

Data ownership often implies that an individual or organization has sole ownership rights. The opposite is the case in the paradigm *everyone as owner*, which is applied when data is intended to be shared with a broad user group. In this case, data ownership is not assigned to any individual or organizational party; instead, everyone can become an owner of certain data, and with the same access rights. This paradigm can relate to open data, which is *"[...] data that anyone can access and use"* (Link et al. 2017). Especially when the data is created in a crowdsourced way – as is the case with OpenStreetMap (OpenStreetMap 2019), for instance – the community is the data owner and everyone shares the same rights to access and use the data, under certain restrictions. Still, open data repositories require data governance, which is often hard to establish when responsibilities are distributed and accountabilities cannot be assigned to an individual or organizational entity. This is especially the case with public health data, but also with data collected in smart cities, for instance. Thus, while open access holds the potential for great innovation, issues develop around privacy, confidentiality, and control of data (Kostkova et al. 2016).

2.3 Approaches to data ownership in the enterprise context

In the enterprise context, data ownership provides the underpinning principles for data governance to define roles, responsibilities, and processes (Loshin 2001; Winter and Meyer 2001). Grover et al. (2018, p. 417) argued that *"without appropriate organizational structures and governance frameworks in place, it is impossible to collect and analyze data across an enterprise and deliver insights to where they are most needed."* The assignment of certain ownership rights to roles has proven to be

beneficial: most importantly, people feel responsible, act in their self-interest, and take care of data. Thus, data ownership has been found to positively impact on data quality and system success (Loshin 2001; Van Alstyne et al. 1995). While the assignment of ownership rights and responsibilities has clear advantages, it can also lead to conflict concerning data sharing (Hart 2002).

Data ownership has been specifically investigated for operational systems (Maxwell 1989; Spirig 1987; Wang et al. 1995) and data warehouses (Winter and Meyer 2001). Operational systems seek to enable business processes with quality data, defined as data that fits its purpose (Wang and Strong 1996). Enterprises have sought to centralize operational systems to ease maintenance and control for IT departments. This has resulted in a misconception that IT departments are the data owner and must be responsible for data quality (Van Alstyne et al. 1995). Business users create the data while executing business processes, but also need high confidence (quality) in the data they use. Thus, in operational systems, it is recommended that data ownership holds to its original aim of ensuring high data quality (Maxwell 1989; Spirig 1987). This implies that the data ownership paradigms *creator as owner* and *consumer as owner* fall together.

While data ownership in operational systems follows the logic of business processes, data warehouses and particularly data marts integrate data from multiple business processes (Watson and Wixom 2007). Data warehouses bring together data from operational systems (*push*). To fulfill a certain information demand (e.g. management report), data is integrated for this particular use in data marts (*pull*). Thus, data ownership in data warehouses and data marts must be data-centric and depends on the number of data integration layers. In the case of one data warehouse and one data mart layer, two ownership types can be distinguished (Winter and Meyer 2001). Since data is typically not changed when it is brought into a data warehouse, data ownership on the data warehouse layer stays the same as in operational systems (*data supply*). On the data mart layer, data is typically changed to fulfill a certain information need. Thus, data ownership on this layer is assigned to the party who requests particular information (*data demand*), which is often also the sponsor of such activities.

2.4 The research gap

Debates about data ownership have multiple facets and, with increasing privacy concerns, they go well beyond the boundaries in which data is created. In the enterprise context, data ownership remains more complex compared to other assets. Still, data ownership is needed to clarify rights and responsibilities to ensure business value with effective data governance (Otto 2011; Tallon et al. 2013). The research distinguishes two approaches to data ownership: In operational systems, data ownership is business process-centric, i.e. the creator and the consumer of operational data are often the same. This stands in contrast to data warehouses, where data ownership is data-centric: the consumer is not the creator, because a data mart integrates data from multiple business processes.

To improve their data capabilities and to create value from BDA, companies create data lakes, in which data is stored without a predefined structure and in raw format, to enable data exploration and innovation (Farid et al. 2016; Madera and Laurent 2016; Watson 2017). This stands in contrast to traditional business intelligence and data warehouse infrastructures, where the structure is predefined and data is cleaned upfront to deliver high-quality reports and insights (Watson 2009). With this paradigm shift, new challenges emerge for enterprises (Grover et al. 2018; Sivarajah et al. 2017): On the one hand, enterprises need to manage much larger volumes and a higher variety of data for multiple purposes (Chen et al. 2012). This imposes higher requirements on data quality, data integration, and data security (Grover et al. 2018). In fact, data quality remains one of the key challenges to enable business value from BDA (Abbasi et al. 2016; Grover et al. 2018; Wamba et al. 2015). On the other hand, the development and operation of analytics go beyond the mere aggregation and visualization of data. With artificial intelligence (AI) (Watson 2017), it is harder to keep track of how data is processed. Further, the high dependency of machine learning applications on data may lead to the risk of high technical debt (Sculley et al. 2015). At the same time, the increasing use of AI is fueling debates about ethical questions. For instance, deep learning techniques operate as *'black box'* algorithms whose

working mechanisms are somehow hard to understand (Castelvecchi 2016). This is why analytics can lead to *"[…] discriminatory effects and privacy infringements"* (Custers 2013, p. 3) and why debates have emerged about accountabilities for algorithmic decision-making (Diakopoulos 2016).

These developments are resulting in new issues and questions relating to data ownership, while showing the relevance of defining accountabilities for data.

3 Methodology

We seek to understand *how enterprises define and adapt data ownership in the BDA context* – a complex phenomenon that requires that one analyze rich information related to the adoption of BDA and the definition of data-related roles in enterprises. This is why we opted for an explorative case study research design, which is well suited for answering *how* questions (Yin 2003) and studying such contemporary phenomena in their particular context (Benbasat et al. 1987; Yin 2003). Specifically, we studied multiple case studies so as to ensure our theory's robustness and to draw generalizable conclusions (Benbasat et al. 1987; Yin 2003).

3.1 Case selection

We integrated our research activities into a research program on data management that included close interactions with 11 data management experts from seven high-profile European companies over 12 months. In early 2019, we initiated an expert group to investigate data management challenges in the context of BDA and met 14 times between January and November 2019. The participants were data experts responsible for establishing organizational and technological structures to manage BDA. They represent large corporations from different industries with some maturity in levering BDA.

Case name	Industry	Size	Key informants	Big data and analytics context
Company A	Fast-moving consumer goods	Revenue: $50B to $100B Employees: ~80 000	Manager: Data governance, Enterprise data architect	Organization: central data and analytics management organization Infrastructure: central big data platform for innovation and industrialization of analytics use cases
Company B	Public transportation and mobility infrastructure	Revenue: $1B to $50B Employees: ~35 000	Leader: Business information management, Data governance manager, Big data platform architect	Organization: central data management organization and central/decentralized data science team Infrastructure: corporate data lake for data exploration/experimentation and the operation of analytics use case
Company C	Manufacturing	Revenue: $1B to $50B Employees: ~90 000	Director: Data architecture and engineering, Project manager: Data platform	Organization: corporate data management organization and central platform team Infrastructure: central data platform to enable digital innovations and scale the operation of data products

Table 2 Selected cases

The discussions in the expert group allowed us to develop an understanding of the current situation and to select three (out of seven) companies for further investigation (see Table 2). These three case companies have established an enterprise data lake as an environment to manage BDA and have introduced data and analytics roles, including the data ownership concept. As each case company has a high BDA maturity and belongs to a different industry, the case selection process followed literal replication logic, leading to similar rather than contrasting results (Benbasat et al. 1987; Yin 2003).

3.2 Data collection

For each enterprise, we identified key informants with strategic and operational responsibility to manage BDA and who are aware of the relevance of and issues relating to data ownership. As starting point, we conducted one initial semi-structured interview with the key informants to understand each's technological and organizational structures to manage BDA. These interviews gave us the opportunity to understand the challenges and approaches concerning assigning accountabilities for data in greater depth. In parallel, we collected primary data through internal documents provided by the firms (e.g. BDA platform designs, role models, and organizational structures). These documents informed us not only about their approach to data ownership, but also about the context and related topics, such as technical infrastructure as well as established roles or processes. Gathering information from multiple sources, including expert interviews and internal documents, allowed for triangulation and ensured construct validity (Yin 2003).

3.3 Within- and cross-case analysis

We performed the case analysis in two steps. First, we conducted a within-case analysis (Yin 2003) to understand the different data ownership types in each enterprise. Here, we used an analysis framework to categorize data ownership types, their descriptions, and the organizational assignment of each type. In a subsequent expert group meeting, we discussed and compared each company's data ownership approach. The discussion helped us to understand the similarities and peculiarities of each case. Second, we performed a cross-case analysis (Yin 2003), comparing the findings of the within-case analysis with one another so as to identify common data ownership types and their responsibilities. Further, we linked each identified type to the corresponding data ownership paradigms suggested by Loshin (2001), which helped us to understand each type's peculiarities in a simplified way. Based on our analysis, we outlined four propositions for data ownership in the BDA context. We discussed our findings in a second expert group meeting, which gave us a better understanding of whether the enterprises agreed with our conclusions or if we had missed aspects we had not reflected on. To verify specific aspects with the case companies and to ensure robust findings, we conducted an additional interview with one key informant from each company.

4 Data ownership in the three case companies

To provide insights into the case setting, we start by presenting the general context, i.e. BDA's role in each enterprise and each's approach to data ownership.

4.1 Company A

Company A is undergoing a digital transformation and is introducing innovative digital products (in addition to its traditional product portfolio), which shifts its core business model from business-to-business to business-to-consumer. Through this change, the company faces an increasing number of data created via sensors embedded in the digital product and in new customer touchpoints (e.g. points of sale or web applications). This data is enabling company A to improve the way it understands and interacts with its customers; but, to lever this data, the company had to enhance its data and analytics capabilities. In a first step, it formed a central group that is responsible for enterprise data and analytics. It also established a data lake as a central big data platform (commercialized Hadoop stack from Cloudera, on-premise and partially in the cloud), which enables data scientists to conduct analytics across the traditional business functions based on internal and external datasets. This platform is primarily used for exploration and experimentation, but also for industrialization of analytics use cases. It has three major components: the data repository for storing and staging data from internal and external sources, data science labs for exploration and experimentation, and data products for industrialization of analytics use cases.

Company A distinguishes three data ownership types: *data source owner*, *platform owner*, and *data product owner*. The *data source owner* is *"primary decision maker about the data entities under his responsibility and accountable for the overall integrity, data lifecycle and data quality of data created in his ownership."* This role is typically assigned at a director level or even above, to the head of a business function that creates but also consumes data of this domain. In the data platform context, the *data source owner "provides approval for data usage in data product."* Thus, company A ensures compliant access to sensitive data (e.g. identifiable personal information). When data is then used in a data product, the company arranges a service-level agreement with the corresponding owner of the data sources so as to ensure quality on both sides. Thus, the *data source owner* must *"fulfill service-level agreements for data products."* The *platform owner* is accountable for the platform infrastructure (*technology stack*) and is assigned to the *head of the digital analytics team*. Concerning data, he *"maintains data sanity and business context while data is going through the technology stack."* This includes that he *"oversees and controls work in data labs."* Further, he *"is accountable for the availability of data pipelines."* In this sense, he must ensure that business requirements for data products are being fulfilled. The *data product owner*, as a head of a business function, represents the data use side and *"addresses business need for data driven by analytics use cases."* This makes him *"accountable for output of the technology stack."* Once a data product is developed and ready to use, he *"ensures the business value of a data product over its lifetime."*

Data owner type	Description	Organizational assignment
Data source owner	*"Primary decision-maker about the data entities under his responsibility and accountable for the overall integrity, data lifecycle and data quality of data created in his ownership."* *"Provides approval for data usage in data product."* *"Fulfils service-level agreements for data products."*	Head of a business function: director level or above
Platform owner	*"Maintains the data sanity and business context while data is going through the technology stack."* *"Oversees and controls work in data labs."* *"He is accountable for the availability of data pipelines."*	Head of the digital analytics team
Data product owner	*"Addresses the business need for data driven by analytics use cases."* *"Accountable for the output of the technology stack."* *"He ensures business value of data product over its lifetime."*	Head of a business function: director level or above

Table 3 Data ownership in case company A

4.2 Company B

Case company B is an infrastructure provider. It is undergoing a digital transformation following a corporation-wide program with three main goals: improve interactions with customers, increase internal efficiency, and enhance capacity management. Thus, the company has invested in new digital applications and sensor technologies to collect data from its assets. Further, it provides noncritical data to third parties through open access so as to stimulate innovation from the outside. Advanced and big data analytics are key drivers of company B's digitalization initiative and are strategically relevant to the company. Thus, it established a central big data platform (commercialized Hadoop stack from Cloudera, on-premise) to provide access to data from diverse sources simultaneously for innovation and production. To ensure the reusability of data on the platform, it was decided that data must be actively managed through corresponding organizational roles and structures. A central data management organization was established to ensure data governance. On the analytics side, a central data science team coordinates the activities, while data scientists form part of each business unit. The platform has four major components: data lake, data labs, data apps, and user homes. The data lake serves as an underlying data storage and processing entity that operates along a staging, an integration, and a business transformation layer. Data labs operate on the data lake and serve the data scientists' need to ex-

plore and experiment with data, for instance, a group of data scientists is accessing machine state data in a data lab to develop a predictive maintenance algorithm. The data app represents an operationalized application that uses data from the data lake, for instance, the predictive maintenance application signals service workers in case of required maintenance activity. A user home comprises specific data from the data lake that is private to the user, for instance, a business analyst conducts ad hoc analyses of daily customers.

Company B distinguishes three of data ownership types on the big data platform, according to its components: *data owner*, *owner of the data lab / data app / user home*, *and owner of the data lake*. The *data owner* is responsible for a data feed in the context of the big data platform and is typically assigned to a business role. Thus, this role is *"responsible for data quality, definition, classification, security, compliance and data lifecycle of a data attribute, set of attributes, or dataset."* The data definition (e.g. documentation in data catalog) and classification must be done when data is brought to the big data platform. This implies that the data owner *"controls reading access to his data through data feed on big data platform and ensures compliant use through the provision of no-join policies under the respect of interests of existing and future data user."* These policies must be revisited as new data is continuously brought to the platform. Since not every data feed has a *data owner* assigned when it is brought to the big data platform, the data user is required to find the *data owner*. If the *data owner* cannot be identified, the user must fill this gap and becomes the owner of the requested data. The *owner of the data lake* is *"accountable for the standardization of the overall big data solution architecture."* This includes that he *"proves the compliance of analytics solutions."* Thus, this role is assigned to the role of the *big data solution architect*, who is also responsible for platform development and provides "*information on planned extensions of the data lake."* This role's responsibilities go beyond the architecture of the big data platform, since he *"ensures that new and valuable data is onboarded to the data lake according to the business need and potential. For this, he searches proactively new data sources, valuates their business potential, and initiates the onboarding process."* In this regard, the *owner of the data lake* serves as a mediator between the *data owner* and the *owner of the data lab / data app / user home*. The latter holds the rights to use data either through a data app that is typically assigned to a business role or through a data lab or user home that is typically assigned to technical roles, for instance, a data scientist. This owner also *"manages access to data lab, app, or user home and is accountable for any activity (operational activity or data privacy) on it over its lifetime."* He is also obliged to inform the *platform owner* about whether the environment still generates value or can be removed. A data scientist, as a user of the *owner of the data lab*, *"needs to comply with a conduct of ethics when working with data in a data lab."*

Data owner type	Description	Organizational assignment
Data owner	*"Responsible for data quality, definition, classification, security, compliance, and data lifecycle of data attribute, set of attributes, or dataset."* *"Controls reading access to his data through data feed on big data platform and ensures compliant use through the provision of no-join policies under the respect of interests of existing and future data users."*	Business role
Owner of the data lake	*"Accountable for the standardization of the overall big data solution architecture. Proves compliance of analytics solutions."* *"Gives information on planned extensions of the data lake."* *"Ensures that new and valuable data is onboarded to the data lake according to the business need and potential. For this, he proactively searches for new data sources, valuates their business potential, and initiates the onboarding process."*	Big data solution architect
Owner of the data lab / data app / user home	*"Manages access to the data lab, app, or user home, and is accountable for any activity (operational activity or data privacy) on it over its lifetime."* *"Data scientists must comply with conduct of ethics when working with data in a data lab."*	Business role for the data app Technical role for the data lab/user home

Table 4 Data ownership in case company B

4.3 Company C

Case company C has a long tradition in the automotive industry. It has invested heavily in R&D to embed software in its products to collect and process data. With this data, the company is seeking to monitor its products' conditions and to provide value adding services to its customers. Thus, it strongly relies on data as an essential component of its future business. For traditional data domains, it has established a corporate organization for master data management. Owing to new requirements to manage sensor data and to develop analytics, company A has extended this function's scope and has set up new organizational units. A central platform team has been built up and manages a platform with a virtualized and physical data lake (Microsoft Azure Cloud) to enable digital innovations and to scale the operation of data products. Company C has also flattened its organizational hierarchies so as to become more agile. Its data platform has two major components: a data hub and data solutions. The data hub connects to the data sources and encompasses a physical and a virtual storage for various types and formats of data. The data solution accesses and processes data to develop/deliver a data application for/to a data consumer.

In the context of the data platform, company C distinguishes between three ownership types: *data domain manager*, *infrastructure owner*, and *data application ownership*. The *data domain manager "controls and monitors the data management for his domain."* Each data domain comprises a homogenous set of data attributes describing a business object, for instance, a customer or an asset. This domain approach to structuring data ownership is a typical approach in organizations with mature data management practices. Company C's *data domain manager "receives requests for data processing and provides data for data usage"* and is accountable for data content and responsible for maintaining data according to business requirements. This role is assigned to a business role in lower management to ensure the efficient handling of requests, which corresponds to company C's agile management approach. Company C does not yet distinguish between the input and output data of a data application. Thus, the *data domain manager* is the owner of input data to the platform and output data of data applications as long as they belong to his domain of responsibility. This includes reporting errors and suggesting improvements. The *infrastructure owner* is accountable for the data platform's development and operation. Thus, he *"oversees the implementation and availability of data pipelines to onboard data to the data hub and provision data to data solutions."* At company C, this role is assigned to the *head of the data platform team*, which is part of the *corporate IT* function. The *business logic owner* is *"accountable for data applications over its lifetime, which includes compliant implementation, the maintenance of data application, and support of users."* This role can either be assigned to a business or/and an IT role (central/decentral) depending on a data application's importance and complexity.

Data owner type	Description	Organizational assignment
Data domain manager	*"Controls and monitors the data management for his domain."* *"Receives requests for data processing and provides data for data usage."* *"Reports errors and suggests improvements."*	Business role: lower management
Infrastructure owner	*"Develops and operates the data platform."* *"Oversees the implementation and availability of data pipelines to onboard data to the data hub and to provision data to data solutions."*	Corporate IT role: Head of the data platform team
Business logic Owner	*"Accountable for a data application over its lifetime, which includes compliant implementation, the maintenance of the data application, and support of users."*	Business or/and IT role: lower management

Table 5 Data ownership in case company C

5 Data ownership types and principles in the context of BDA

Through a cross-case analysis, our study has unveiled significant changes and extensions to data ownership with BDA. Three ownership types were present in all three enterprises: *data owner*, *data platform owner*, and *data product owner* (see Table 6).

Proposition 1: In the context of BDA, companies define data ownership at three levels: data source or dataset (data supply), data product (data demand), and data platform.

We will now present and discuss each data ownership type and will link it to the corresponding data ownership paradigm suggested by Loshin (2001). This link helps us to understand the peculiarities of each type in a simplified way and the related phisolophical assumptions. We will then formulate propositions for data ownership in the context of BDA, synthesizing key implications and requirements to manage BDA.

The *data owner* is first the creator but can also be user of data (sources) in his or her domain of responsibility. This implies the accountability for the quality and the lifecycle of data, and can be associated with the paradigm of *creator as owner.* This is a very important role in data organizations, since data quality remains one of the key challenges to enable business value from BDA (Abbasi et al. 2016; Grover et al. 2018; Wamba et al. 2015). The *data owner* is a pure business role in all three case companies, but with varying organizational assignment levels. While in company A, this role is assigned on a director level, in company C, it is assigned to a lower management function so as to ensure efficiency in handling data requests. We have demonstrated that BDA extends the responsibilities of *data owners* to also provide the input data for new data products. First, the data owner is expected to address the particular requirements of data products according to service-level agreements – as in company A. Second, the *data owner* ensures compliant access and use of the data on the platform, i.e. manages data requests, approves usage, and provides access. For instance, the *data owner* in company B must continually revisit the no-join policies so as to ensure compliant use, also when the number of data available on the platform increases. This requires both additional effort and knowledge of potential implications when data is combined with data from other domains. In this regard, the *data owner* controls the decentralized access, which is one of the key data security issues to be solved in BDA environments (Grover et al. 2018), and may even be needed at an intra-organizational level (Günther et al. 2017).

Proposition 2: The data owner ensures compliant access to and use of data, not only in the source system, but also on the platform and in data products. This extends beyond the traditional scope of responsibility and requires one to manage more data dependencies.

The *data product owner* is accountable for the data product. Notably, the companies differentiated between data products that provide access to data for exploration and experimentation purposes (typically, a data lab for data scientists) and data products in production. In the latter case, company A defined this role to mainly ensure that the data product generates a business value over its lifetime. In this sense, the *data product owner* can be linked to the *consumer as owner* paradigm. However, other interpretations are possible for the *data product owner*. In case companies A and C, the *data product owner* is accountable for the data product over its lifetime, including development, maintenance, and user support. Here, the paradigms *decoder as owner* (e.g. a data scientist who decodes a pattern in the data) or *compiler as owner* (e.g. data analysts who aggregate multiple data sources) are more suitable as the *data product owner* involved in the creation of the data product that is then consumed by a user.

In the BDA context, data products are getting more complex than in the traditional data warehouse context, where the data mart layer is mostly owned by the same business function as the data source. The consumer of this data is then also the owner of the data product. This split also aligns with the principle in data warehouse systems where the creator of data (e.g. a local sales manager) is a different owner to the consumer of certain information (e.g. a head of global sales) (Winter and Meyer 2001). But while the purpose of data in the data warehouse is known, in the BDA context, its purpose is unknown. This makes controlling difficult.

Proposition 3: The data product owner ensures business value of a data product over its lifetime, including maintenance and user support. Depending on the data product's complexity, this role may require technical expertise; thus, this may be a shared role between business and IT.

Data owner type	Responsibilities	Support in cases	Exemplary statement
Data owner	Accountable for quality and lifecycle of data in his domain of responsibility.	A, B, C	*"[…] accountable for the overall integrity, data lifecycle, and data quality of data created in his ownership."* (A)
	Fulfils quality requirements for data in his domain of responsibility for data products.	A	*"Fulfils service-level agreements for data products."* (A)
	Ensures compliant access and use of data in his domain of responsibility by handling requests, providing access, and approving usage.	A, B, C	*"Controls reading access […] ensures compliant use through the provision of no-join policies […]."* (B)
Data platform owner	Ensures data quality on the platform by managing data pipelines to onboard and provision data.	A, C	*"Oversees the implementation and availability of data pipelines to onboard data to the data hub and to provision data to data solutions."* (C)
	Accountable for onboarding of valuable data according to a business need and potential.	B	*"Ensures that new and valuable data is onboarded to the data lake according to the business need and potential."* (B)
	Responsible for the development and operation of the data platform. Approves compliance of data products according to data platform standards.	B, C	*"Develops and operates the data platform."* (C)
Data product owner	Ensures that a data product addresses a business need and generates business value over its lifetime.	A	*"He ensures business value of a data product over its lifetime."* (A)
	Accountable for a data product over its lifetime, including development, maintenance, and user support.	A, C	*"Accountable for a data application over its lifetime, which includes compliant implementation, maintenance of the data application, and support of users."* (C)
	Ensures compliant access and use of data product.	B	*"Manages access to data lab, app, or user home and is accountable for any activity […] on it over its lifetime."* (B)

Table 6 Data ownership types in the context of big data and analytics

Companies manage BDA with data platforms, storing data from multiple sources and delivering data products for data exploration/experimentation and for direct use. This observation underpins the disruptive nature of BDA to amalgamate technologies to derive knowledge from big data into platforms (Abbasi et al. 2016). All enterprises have the role of a *data platform owner,* which serves as a mediator and facilitates data supply (*data owner*) and data demand (*data product owner*). While there are many *data owners* and *data product owners*, there is usually only one *data platform owner* assigned to an IT role in an enterprise. Thus, we can link this ownership type to the paradigms *compiler as owner*, since this role brings data from various sources to the platform, and *packager as owner*, since they reformat data for particular uses in data products. In company B, this role has the important (even strategic) function to *"proactively"* search for and bring valuable data (according to a business potential and need) to the platform. This role is also accountable for the development and operation of the platform – as is also the case in company C. This also includes controlling whether data products comply with data platform standards. In sum, the *data platform owner* is responsible for the availability of data on the platform, since he or she manages the data pipelines to bring data to the platform and to provide data to data products. Our findings thereby also support Wamba et al.'s (2015, p. 242) study that *"[…] emphasizes not only the support but also the active involvement of senior management for successful implementation of the shared platform to leverage 'big data' capabilities."*

Proposition 4: In BDA environments, the data platform owner role facilitates data supply (data owners) and data demand (data product owners). This ensures the availability of data on the platform for data exploration and experimentation, but also the operation of data products.

6 Conclusion and outlook

6.1 Findings and limitations

Our findings confirm that data ownership remains a key concept to clarify rights and responsibilities, but should be revisited in the BDA context. Some of the established principles for operational systems and data warehouses still hold true; most importantly, the clear distinction between the owner on the data supply side (*data owner*) and the owner on the data demand side (*data product owner*). Thus, the *data owner* is accountable for data as the input to data products, and the *data product owner* ensures the business value from the data product. This is similar to the owner of the data warehouse, who provides data, and the owner of the data mart, who outlines an information need (Winter and Meyer 2001). Despite these similarities, BDA environments require also a change in responsibilities. In a data warehouse environment, access provisioning and data quality requirements are more predictable and can be clarified at the outset. The opposite is true for the BDA environment (data platform), where the purpose of data is intentionally unknown, to allow for data exploration and experimentation. This implies that the *data owner* must manage data access and must continually react to changing data requirements. Further, data on the platform is usually freely accessible, which holds risks for compliant data use, especially when it can be combined with data from other domains. This requires knowledge from *data owners*, which goes beyond their domain of expertise. The *data product owner* is accountable for the business value of BDA applications over their lifetimes. This also requires technical expertise, since data processing and analysis is becoming more complex with BDA, for instance, through machine learning components (Sculley et al. 2015). Thus, this role may be shared between a business and an IT role. This is also why we observed that companies define the *data platform owner* is required to mediate data supply (*data owner*) and data demand (*data products*).

This study has certain limitations. Since the three case companies represent large organizations, the findings may not be transferrable to smaller enterprises. Also, case studies only allow for analytical generalization, and we suggest quantitative empirical studies to further validate our findings.

6.2 Theoretical and practical implications

Our research has provided fundamental considerations and empirical insights around data ownership in the BDA context, with implications for practice and theory. As data ownership helps one to clarify rights and responsibilities, the identified data ownership principles and types can form the basis for more comprehensive data governance roles and frameworks.

Practitioners may use our findings in the context of data governance initiatives to define their approach to ownership as well as the related roles and responsibilities. For researchers, our study lays the theoretical foundations for effective management and of organizational roles for BDA. It links data ownership to the general philosophical assumptions and also opens multiple avenues for future research: First, our propositions and the suggested ownership types represent a first step towards studying *BDA governance to facilitate the value creation process,* which is a key theme of Grover et al.'s (2018) research framework. Further, the ownership types and governance structures need to be complemented by new approaches to data quality management that are required to enable data exploration and experimentation, in combination with processes to ensure efficient data onboarding to the platform and data product delivery.

References

Abbasi, A., Sarker, S., and Chiang, R. 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems* (17:2). (http://aisel.aisnet.org/jais/vol17/iss2/3).

Alexander, D., and Lyytinen, K. 2017. "Organizing Successfully for Big Data to Transform Organizations," *AMCIS 2017 Proceedings*. (http://aisel.aisnet.org/amcis2017/DataScience/Presentations/30).

Benbasat, I., Goldstein, D. K., and Mead, M. 1987. "The Case Research Strategy in Studies of Information Systems," *MIS Quarterly*, pp. 369–386.

Castelvecchi, D. 2016. "Can We Open the Black Box of AI?," *Nature News* (538:7623), p. 20. (https://doi.org/10.1038/538020a).

Chen, H., Chiang, R. H., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact.," *MIS Quarterly* (36:4), pp. 1165–1188.

Chessell, M., Scheepers, F., Strelchuk, M., Starre, R. van der, Dobrin, S., and Hernandez, D. 2018. "The Journey Continues: From Data Lake to Data-Driven Organization," Redbooks.

Comuzzi, M., and Patel, A. 2016. "How Organisations Leverage Big Data: A Maturity Model," *Industrial Management & Data Systems* (116:8), pp. 1468–1492. (https://doi.org/10.1108/IMDS-12-2015-0495).

Confessore, N. 2018. "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far," *The New York Times*. (https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html).

Custers, B. 2013. "Data Dilemmas in the Information Society: Introduction and Overview," in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Studies in Applied Philosophy, Epistemology and Rational Ethics, B. Custers, T. Calders, B. Schermer, and T. Zarsky (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3–26. (https://doi.org/10.1007/978-3-642-30487-3_1).

Davenport, T. H., Barth, P., and Bean, R. 2012. "How 'Big Data' Is Different," *MIT Sloan Management Review* (54:1), p. 5.

Diakopoulos, N. 2016. "Accountability in Algorithmic Decision Making," *Communications of the ACM* (59:2), pp. 56–62. (https://doi.org/10.1145/2844110).

Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F., and Chu, X. 2016. *CLAMS: Bringing Quality to Data Lakes*, ACM Press, pp. 2089–2092. (https://doi.org/10.1145/2882903.2899391).

Grover, V., Chiang, R. H. L., Liang, T.-P., and Zhang, D. 2018. "Creating Strategic Business Value from Big Data Analytics: A Research Framework," *Journal of Management Information Systems* (35:2), pp. 388–423. (https://doi.org/10.1080/07421222.2018.1451951).

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., and Feldberg, F. 2017. "Debating Big Data: A Literature Review on Realizing Value from Big Data," *The Journal of Strategic Information Systems* (26:3), pp. 191–209. (https://doi.org/10.1016/j.jsis.2017.07.003).

Hart, D. 2002. "Ownership as an Issue in Data and Information Sharing: A Philosophically Based Review," *Australasian Journal of Information Systems* (10:1). (https://doi.org/10.3127/ajis.v10i1.440).

Janeček, V. 2018. "Ownership of Personal Data in the Internet of Things," *Computer Law & Security Review* (34:5), pp. 1039–1052. (https://doi.org/10.1016/j.clsr.2018.04.007).

Jones, C., and Tonetti, C. 2019. "Nonrivalry and the Economics of Data," No. w26260, Cambridge, MA: National Bureau of Economic Research, September. (https://doi.org/10.3386/w26260).

Kostkova, P., Brewer, H., de Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., Koczan, P., Knight, P., Marsolier, C., McKendry, R. A., Ross, E., Sasse, A., Sullivan, R., Chaytor, S., Stevenson, O., Velho, R., and Tooke, J. 2016. "Who Owns the Data? Open Data for Healthcare," *Frontiers in Public Health* (4). (https://doi.org/10.3389/fpubh.2016.00007).

Labadie, C., and Legner, C. 2019. "Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR," *Wirtschaftsinformatik 2019 Proceedings*. (https://aisel.aisnet.org/wi2019/track11/papers/3).

Link, G., Lumbard, K., Germonprez, M., Conboy, K., and Feller, J. 2017. "Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations," *Communications of the Association for Information Systems* (41), pp. 587–610. (https://doi.org/10.17705/1CAIS.04125).

Loshin, D. 2001. *Enterprise Knowledge Management: The Data Quality Approach*, Morgan Kaufmann.

Madera, C., and Laurent, A. 2016. "The Next Information Architecture Evolution: The Data Lake Wave," in *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, MEDES, New York, NY, USA: ACM, pp. 174–180. (https://doi.org/10.1145/3012071.3012077).

Maxwell, B. 1989. "Beyond 'Data Validity': Improving the Quality of HRIS Data," *Personnel* (66:4).

McKenna, E., Richardson, I., and Thomson, M. 2012. *Smart Meter Data: Balancing Consumer Privacy Concerns with Legitimate Applications*. (https://doi.org/10.1016/j.enpol.2011.11.049).

OpenStreetMap. 2019. "OpenStreetMap," *OpenStreetMap*. (https://www.openstreetmap.org/copyright, accessed November 13, 2019).

Otto, B. 2011. "Data Governance," *Business & Information Systems Engineering* (3:4), pp. 241–244. (https://doi.org/10.1007/s12599-011-0162-8).

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. 2015. "Hidden Technical Debt in Machine Learning Systems," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), Montreal, Canada, pp. 2503–2511.

Shleifer, A. 1998. "State versus Private Ownership," *Journal of Economic Perspectives* (12:4), pp. 133–150.

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. 2017. "Critical Analysis of Big Data Challenges and Analytical Methods," *Journal of Business Research* (70), pp. 263–286. (https://doi.org/10.1016/j.jbusres.2016.08.001).

Spirig, J. 1987. "Compensation: The up-Front Issues of Payroll and HRIS Interface," *Personnel* (66:100), pp. 124–129.

Tallon, P. P., Ramirez, R. V., and Short, J. E. 2013. "The Information Artifact in IT Governance: Toward a Theory of Information Governance," *Journal of Management Information Systems* (30:3), pp. 141–178. (https://doi.org/10.2753/MIS0742-1222300306).

The Economist. 2017. "Data Is Giving Rise to a New Economy," *The Economist Group Limited*. (https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy).

Van Alstyne, M., Brynjolfsson, E., and Madnick, S. 1995. "Why Not One Big Database? Principles for Data Ownership," *Decision Support Systems* (15:4), pp. 267–284. (https://doi.org/10.1016/0167-9236(94)00042-4).

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D. 2015. "How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study," *International Journal of Production Economics* (165), pp. 234–246. (https://doi.org/10.1016/j.ijpe.2014.12.031).

Wang, R. Y., Storey, V. C., and Firth, C. P. 1995. "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering* (7:4), pp. 623–640. (https://doi.org/10.1109/69.404034).

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5–33.

Watson, H. 2009. "Business Intelligence: Past, Present and Future," in *AMCIS 2009 Proceedings*, San Francisco.

Watson, H. J. 2017. "Preparing for the Cognitive Generation of Decision Support," *MIS Quarterly* (16:3).

Watson, H. J., and Wixom, B. H. 2007. "The Current State of Business Intelligence," *Computer* (40:9), pp. 96–99. (https://doi.org/10.1109/MC.2007.331).

Winter, R., and Meyer, M. 2001. "Organization of Data Warehousing in Large Service Companies - A Matrix Approach Based on Data Ownership and Competence Centers," *Journal of Data Warehousing* (6:4), pp. 23–29.

Wixom, B., and Ross, J. 2017. "How to Monetize Your Data," *MIT Sloan Management Review* (58:3), p. 7.

Yin, R. 2003. *Case Study Research: Design and Methods, Third Edition, Applied Social Research Methods Series, Vol 5*, London, UK: Sage Publications, Inc.