Creating a Data Factory for Data Products

Chris Schlueter Langdon[1(✉)] and Riyaz Sikora[2]

[1] Peter Drucker School of Management, Claremont Graduate University, 1021 N Dartmouth Ave Claremont, Claremont, CA 91711, USA
chris.langdon@cgu.edu

[2] Information Systems, College of Business, University of Texas at Arlington, P.O. Box 19437, Arlington, TX 76019, USA
rsikora@uta.edu

Abstract. Data is seen as the next big business opportunity. From a demand side, the popularity of artificial intelligence (AI) is growing and particularly deep learning requires large amounts of data. From a supply side, new technology, such as Internet of Things (IoT) sensors and 5G mobile communications, will greatly expand data generation. However, data has remained a challenge. In data analytics companies are struggling with too much time spent on data preparation. As of today, data preparation for analytics has largely remained handmade and made-to-order like cars before Henry Ford industrialized the auto business through productization of cars and parts, and factory automation. Similarly, for data analytics to become a bigger business, data has to be productized. First "data factories" are emerging to create such data products economically. This article introduces a framework to guide construction of a data factory: What are the key modules, why are they important, how is best practice evolving? The article is building on (a) a foundation and in-depth case studies in the literature, (b) current meta research and systematic literature reviews (SLRs), and (c) our own observations building a data factory. This real-world application uncovered the important additional steps of data rights management and data governance that may be less obvious from a computer science perspective but critically important from a business and information systems view.

Keywords: Data product · Data factory · Data sovereignty · Data governance · Data quality

1 Data as the Next Big Business

Data is promised to be the next big business (e.g., Wall 2019, Gartner 2018a). Investment banks, analysts and consultants further feed the frenzy with big revenue forecasts. In terms of data monetization opportunities, consultants McKinsey & Company estimate that car-generated data alone will be worth between US$450 billion and US$750 billion by 2030, less than two vehicle generations away (McKinsey 2016). Consumer data is already a business today. Google and Facebook live off the data that users create on their platforms. Almost all their revenue is from advertising, selling "eyeballs" and user engagement to advertisers.

K. R. Lang et al. (Eds.): WeB 2019, LNBIP 403, pp. 43–55, 2020.
https://doi.org/10.1007/978-3-030-67781-7_5

Lesser known consumer data companies are market researchers, like GfK, Ipsos and Nielsen - the top marketing research companies according to the (American Marketing Association 2018). Yet, the list of data vendors is far longer. A new 2019 Vermont law requires data brokers to be registered (Vermont 2018), and already a list of more than 120 companies has emerged (Melendez 2019).

All of the above is just the beginning. A big data boost is expected from IoT data (Internet of Things): IoT is essentially turning objects into websites. Historically, the Web and website tracking created a first wave of Big Data (which in turn created new technology to store and process it, such as Hadoop). Now ordinary objects are turned in to websites. For example, cars: connected and autonomous vehicles are projected to generate four terabytes (TB) of data a day (Krzanich 2016). Furthermore, this IoT boom is fueled by a confluence of trends in information systems, such as miniaturization of sensors like lidar (light detection and ranging sensor for autonomous cars), device technology, e.g., edge computing, and a new 5G cellular mobile communications standard.

1.1 The Problem: Data Isn't Scaling

A key mechanism to release value from data is analytics. With Websites it took tools like Google Analytics (Urchin) to benefit from Website tracking and attract advertising budgets. Google Analytics is mostly descriptive analytics. Far more value is generated from consecutive stages of predictive and prescriptive analytics (McKinsey 2018, Gartner 2018b). Examples include product recommendations using machine learning as an amplifier of word-of-mouth marketing (e.g., on Amazon and Netflix; see also Stern et al. 2009); and the application of deep learning or neural network methods across many domains for the recognition of text (sentiment analysis), picture (automatic license plate recognition, ALPR), video (autonomous vehicles) and speech recognition (Amazon's Alexa virtual assistant). Yet despite the media hype a quick review of time spent in data analytics projects reveals a big problem. Today, according to the literature more than 80% of the time budget of a data analytics project is spent on data wrangling - not with algorithms (Press 2016, Vollenweider 2016). Companies have gone from databases to data warehouses and now to data lakes (Porter and Heppelmann 2015) - and they seem to be drowning in it. Our own survey of data experts confirms the problem. If an analytics project is broken into the three phases of (a) data processing, (b) analytics modeling & evaluation, and (c) deployment, then timeshares are reported as 48%, 32% and 20% respectively ($n = 65$, our survey has been conducted in 2018 using a convenience sample of data experts at data science events for business – not academic conferences).

1.2 The Solution: Data Productization

These numbers confirm that data processing for AI remains handmade just like cars before Henry Ford industrialized auto making. Gottlieb Daimler invented the motor car in 1886, but it was Henry Ford who invented the modern auto business about 20 years later (Womak et al. 1990). He evolved auto making from a hand-made affair to mass production through automation, which made autos affordable for a big market. The moving assembly line is probably the most visible and striking feature. However,

less obvious, for automation to work Ford critically required interchangeability of parts, which in turn required metrics (Clark and Fujimoto 1991). Parts had to be made to precise measurements so that all copies of a part were more or less similar in order to be attached to cars coming down the line quickly without lengthy calibration and refitting work. Mechanical engineering introduced the notion of tolerance as "the range of variation permitted in maintaining a specified dimension in machining a piece" (Webster 2019). Parts were specified ("specced") in engineering drawings or "blueprints" and then manufactured within precise tolerances to make them interchangeable.

So far, data has eluded proper measurements and is in need of productization (Crosby and Schlueter Langdon 2019, Glassberg Sands 2018). Data attributes have remained qualitative and subjective. Examples include fundamental properties, such as measures of size and quality. How to size data? What is big data? Is size measured in (i) bytes or (ii) population size or (iii) length of a time series - or all of the above? Our survey affirms the complication. All three dimensions seem to matter (bytes: 24%, population size: 31%, length of time series: 45%, n = 67). Same with quality: Without metrics data remains ambiguous like parts that may or may not fit, which will also inhibit data sharing and exchange. Akerlof has demonstrated how the lack of transparency of attributes or "asymmetry of information" between buyers and sellers will lower product quality (lemons) or prevent market exchange outright (Akerlof 1970).

2 Productization in "Data Factories"

In 2006 Clive Humby, a mathematician and architect of UK retailer Tesco's club and loyalty card, spoke of "data as the new oil" at the Association of National Advertisers' marketer's summit at Chicago's Kellogg School of Management (ANA 2006). The one part of his oil analogy, that data may be as valuable as oil has caught on - although data is not even used up in consumption like oil. The other part about the refining effort has not. Humby's analogy suggests that in order to prepare raw data into a refined data product for analytics applications - AI-ready data - it may take extensive refining and at an industrial scale with large platforms comparable to massive refineries for oil.

IT industrialization is certainly not a new phenomenon (Walter et al. 2007). And for plain data storage and processing this refinery analogy appears to correspond very well with observations in the field, specifically the explosive growth of the cloud business.

Launched with Amazon's Elastic Compute Cloud in 2006 the global public cloud service revenue for 2019 has been estimated to exceed US$200 billion (Gartner 2018b). Furthermore, supporting Humby's scale argument, the business is already highly concentrated at an early age with only three hyperscalers dominating most of the business: Amazon's Web Services (AWS), Microsoft's Azure, and Google's Cloud Platform (GCP). As of end of 2018 these top 3 vendors accounted for 60% of the business, the top 10 for nearly 75% (Miller 2019).

2.1 Toward a Data Factory Framework

The conceptualization of our data factory framework builds on an established foundation. It has evolved in a multi-step investigation from (a) in-depth case study analysis in

the literature and (b) systematic literature reviews (SLRs) to (c) our own observations building a data factory in practice. Figure 1 summarizes developments in the literature as the foundation of our refinements.

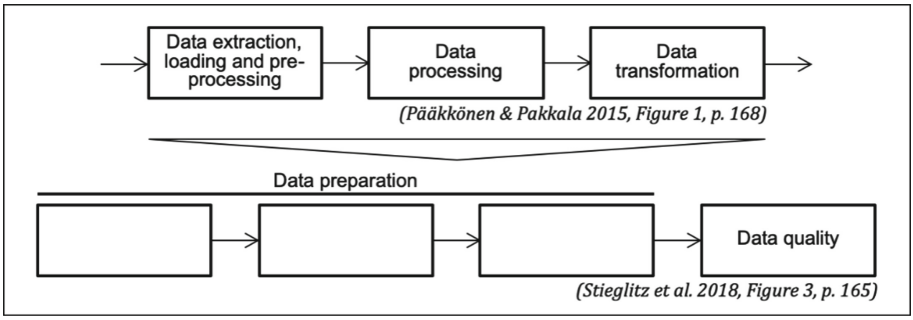

Fig. 1. Evolution of framework foundation in the literature

Pääkkönen & Pakkala present a first analysis of internal "data factories" using in-depth case studies of big data pioneers (2015). The authors dissect data operations at pioneers like Facebook and Netflix and establish that data preparation at these companies is a 'process' as "a series of actions or steps" (Webster) analogous to a 'factory' as "a set of […] facilities for […] making wares […] by machinery" (Webster). This stepwise decomposition conforms with the evolution of information system capabilities toward modularization and flexibility as seen with the emergence of Web services, for example with Microsoft's .NET framework (Schlueter Langdon 2006, 2003b). Specifically, Pääkkönen & Pakkala reveal three major and common steps of data refinement – because of our focus on data refinement, we are explicitly excluding any analysis, analytics and visualization steps: (i) data extraction, loading and pre-processing; (ii) data processing, and (iii) data transformation. This in-depth, case-study based assessment of big data pioneers is corroborated through extensive SLRs: The first study includes 227 articles from peer-reviewed journals extracted from the Scopus database from 1996–2015 (Sivarajah et al. 2017).

It confirms three steps in the data preparation process (again, excluding data analysis, analytics and visualization steps): data intake (acquisition and warehousing), processing (cleansing) and transformation (aggregation and integration; p. 273).

A second, recent study considered 49 articles from three different branches in the literature (Stieglitz et al. 2018): computer science (ACM and IEEE), information systems (AIS), and the social sciences (ScienceDirect). This second SLR yields the addition of data quality as another distinct and common step in the data refinement process (Stieglitz et al. 2018, Fig. 3, p. 165). These four steps as illustrated in Fig. 1 provide the foundation to which we add our observations constructing a real-world data factory. This factory is built by Deutsche Telekom and part of the Telekom Data Intelligence Hub (DIH, Deutsche Telekom 2018). Deutsche Telekom is one of the world's leading integrated telecommunications companies, with some 178 million mobile customers and nearly 50 million fixed-network lines; it operates in more than 50 countries, and generated

revenue of 76 billion Euros in the 2018 financial year (Deutsche Telekom 2019). The DIH has been launched as a minimum viable product in late 2018 in Germany at: https://dih.telekom.net/en/. Coming from this practical experience we propose a slightly more granular decomposition of data refinement activities to explicitly recognize issues that have emerged as a critical concern in practice and that require additional data processing steps: data privacy and data sovereignty. Both issues had already surfaced in the SLR by Sivarajah et al. but only as "management challenges" not explicitly as data refinement steps (p. 274). However, since 2018, the General Data Protection Regulation (GDPR) is mandating data privacy protection in the entire European Union, which necessitates additional data refinement steps, such as consent management, anonymization and user data deletion (European Commission 2018). Similarly, the issue of data sovereignty has evolved from a hygiene factor to a key element of a company's business strategy (e.g., Otto 2011) – it even factors into industrial policy of nations: "the question of data sovereignty is key for our competitiveness,' according to Germany's economy minister" (Sorge 2019). And Europe is not alone; in 2018 California became the first U.S. state with a comprehensive consumer privacy law when it enacted the California Consumer Privacy Act of 2018 (CCPA), which becomes effective 2020 (Cal. Civ. Code §§ 1798.100-1798.199). CCPA not only grants residents in California new rights regarding their personal information; more importantly it imposes data protection duties on entities conducting business in California. This matters, because California is a very big market. It is the most populous state in the U.S. and based on its GDP it would rank as the fifth largest economy in the world ahead of Great Britain, France and Italy Fig. 2 (http://www.dof.ca.gov/Forecasting/Economics/Indicators/Gross_State_Product/).

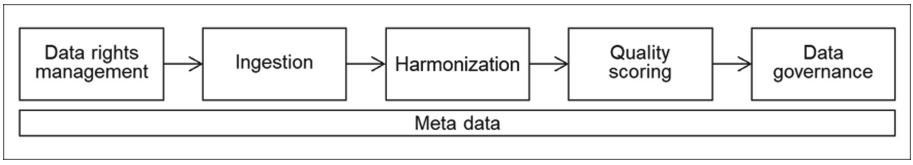

Fig. 2. Extended data factory framework

Legal issues may not be so important from a pure computer science and software engineering perspective. For information systems they certainly matter, because any information system and its architecture would have to correspond with business requirements (Schlueter Langdon 2003a). Therefore, we propose to bookend the data refinement process by data rights management at the beginning to ensure any refinement is compliant with legal requirements in the first place and by data governance at the end to safeguard data sovereignty. Figure 2 illustrates the expanded data factory framework.

In a nutshell raw data rights must be verified before any data can be ingested or harvested (rights, licensing, user consent). Then data ought to be harmonized or properly labeled or tagged for it to be made discoverable through a catalog of categories and search engines (classification). Furthermore, it needs to be scored to provide some indication of quality, because without it any subsequent analytics is pointless – "garbage in, garbage out" (GIGO, quality scoring). Finally, governance mechanisms are required

to ensure that data can be exchanged while data sovereignty is maintained for each data provider. For example, in early 2019, Telekom DIH became the first platform to offer data governance controls based on an architecture developed by a consortium of Fraunhofer institutes as illustrated in Fig. 3 (Fraunhofer 2019). Other data factories are emerging. Microsoft is offering "Azure Data Factory" as a feature in its Azure cloud, which is alleviating fears in Europe that hyperscalers are already expanding their dominance beyond data storage (Clemons et al. 2019). In the Azure Data Factory users can "create and schedule data-driven workflows (called pipelines) that can ingest data from disparate … [sources and]… move the data as needed to a centralized location for subsequent processing" (Microsoft 2018). A quick comparison of this description with Fig. 2 reveals that it is so far more narrowly focused on an upstream module, specifically on ingestion. Another "Data Factory," by Datahub, provides open toolkits for data cleaning, modification and validation (Datahub 2019), which - according to Fig. 2 - would be more focused downstream on data classification and quality enhancements.

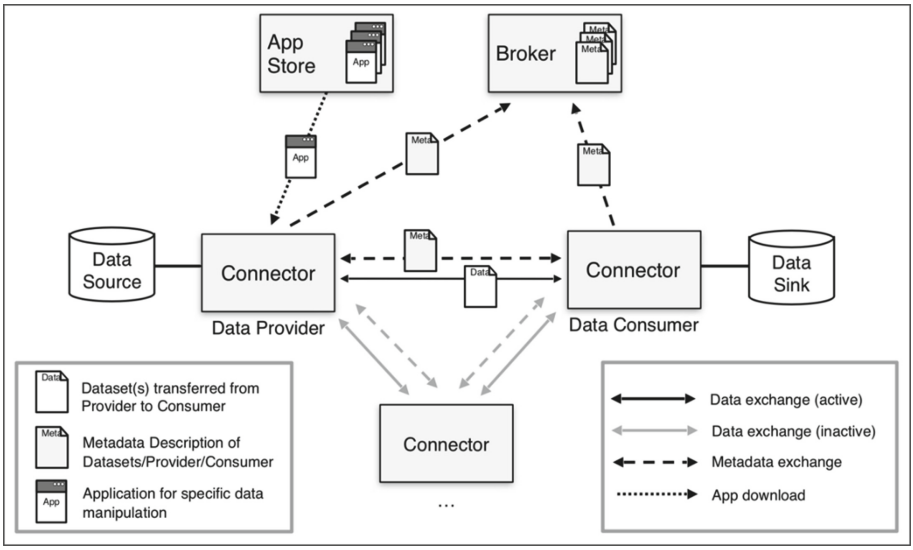

Fig. 3. Data governance architecture (IDSA 2019, p. 59)

2.2 Rights Management

For any product to be marketed and sold ownership rights and licensing rules have to be verified and observed. With data the issue of ownership and rights is complicated. In Germany for example, ownership is typically limited to physical objects ("Sachen, körperliche Gegenstände," § 903 German Civil Code BGB, BMJV 2013). Electronic data is not included. Instead data is protected and rights to it are dealt with in special regulation, such as data protection, copyright and competition laws (for an overview, see Dewenter and Lueth 2018, Chapters 4 and 5). However, since May 2018 complication

with personal data rights in the European Union (EU) has been greatly reduced. The EU is now enforcing the General Data Protection Regulation (GDPR, European Commission 2018). GDPR aims to give control to individuals in the EU over their personal data. It has created the legal foundations for a uniform digital single market, which greatly simplifies the regulatory environment for doing business in the EU.

It also has legitimized the business that involves personal data including the user-product interaction data or behavioral data, which has proven to be most valuable for optimization, customization and personalization of customer journeys and the user experience (UX; Crosby and Schlueter Langdon 2014). In a nutshell, in EU countries, any personal data has to be GDPR compliant.

Key requirements include user consent, privacy, fair and transparent processing, limits to storage and use, right to be forgotten, user access and portability and breach reporting duties. GDPR gives individuals control over their personal data and privacy. Unless an individual has provided explicit, informed consent to data processing, personal data may not be processed. Instead of stifling data innovation GDPR has an opposite effect, igniting data industrialization because it legitimizes the use of personal data for business. Being GDPR compliant can avoid embarrassment, erosion of trust and legal trouble. Particularly data leaks involving social media data have accumulated to trigger a public backlash ("surveillance capitalism," Zuboff 2019) to the point that Facebook's CEO Zuckerberg is encouraging introduction of regulation similar to GDPR in the US (Zuckerberg 2019).

For any data factory a first key step will include data rights management to ensure compliance with licensing agreements and GDPR. In case of personal data or data with personally identifiable information an important step is anonymization and pseudonymization. According to Recital 26 of GDPR anonymized data must be stripped of any identifiable information, so that it becomes impossible to generate insights on a discreet individual, even by the party that is responsible for the anonymization. Pseudonymization is less restrictive and requires personal data to be separated into usable data and "additional information" so that "data can no longer be attributed to a specific data subject without use of additional information" (Article 4(5) GDPR). If a data factory is processing particularly personal data at a large scale, a Data Protection Officer (DPO) must be appointed (Article 37 GDPR).

2.3 Ingestion

An early challenge in the data production process is to retrieve all relevant data for an AI application or data product. Often, it involves connecting all required sources and moving the data to a centralized location for subsequent processing (examples of Facebook and Netflix in Pääkkönen and Pakkala 2015). This can become a cumbersome affair as data is typically scattered throughout an enterprise and its supply chain and channel system. In most companies data is stashed away in databases, lots of databases - reflected in the growth of database vendors, like Oracle. Some data has been moved into data warehouses and lately into data lakes (Porter and Heppelmann 2015) - some is held on premise; other data is already stored in public clouds.

In addition to locating and retrieving data from multiple sources it may exist in different formats - and may have to be converted or transformed. For each data type

(text, image, audio, video) there are multiple file format options. Examples range from plain text (txt) and csv (comma-separated value) files to video formats (such as MPEG-4), and from open-standard file formats, such as JSON, to proprietary formats (USGS 2019, Oregon State University 2019).

Depending on the application domain and analytics method data may have to be transformed into a form consumable by a particular AI method. This may require additional activities, such as normalizing it and dealing with missing data, corrupted binary data and miss-labeled column descriptions.

2.4 Harmonization

The Cambridge Dictionary defined harmonization as "the act of making different [… elements] suitable for each other, or the result of this (Cambridge Dictionary). For data this includes data classification, which is important for at least two uses cases: (1) discovery and (2) training. In order to reduce the time it takes to find the right information, data has to be labeled or tagged so it can be discovered quickly either through a catalog or search engine. Classification is also required for training data. Raw data has to be labeled and annotated for use in training and validation of machine learning systems. For example, in autonomous driving with footage from onboard cameras, someone must go through each frame and identify people, objects and markings. The deep learning system needs to be told what pedestrians look like and from different angles in different weather conditions in order to generate the computational model of that pattern.

These additional information on raw data is referred to as metadata, which is created along the entire data production pipeline (see Fig. 2). One issue with metadata and labeling in particular is semantic standardization. Different areas and companies introduce their schematic and principles or "vocabulary," which may or may not be compatible with other vendors or domains. These schematics and underlying principles are also referred to as data taxonomies (taxonomy comes from the Greek τάξις, taxis - meaning 'order', 'arrangement'; and νόμος, nomos - 'law' or 'science'; Merriam-Webster).

For example, in areas such as autonomous driving, many specialists have emerged that focus on labeling and annotating raw data for use in training AI. This group includes well established vendors of high definition maps, such as Here and TomTom, as well as startups eager to create their own taxonomy as intellectual property. Recently, two open harmonization efforts have gained traction: OpenDrive to describe entire road networks with respect to all data belonging to the road environment, and OpenScenario to describe the entities acting on or interacting with the road.

On the Web, the The World Wide Web Consortium (W3C), the international community that develops open Web standards, has published a framework and recommendations on web annotation complete with a model (describes underlying abstract data structure), vocabulary (which underpins the model) and protocol (HTTP API for publishing, syndicating, and distributing Web Annotations) (https://www.w3.org/annotation/).

2.5 Quality Scoring

Quality scoring is a well-established business - but not with data. As consumers, most of us are probably familiar with Consumer Reports in the US ("Stiftung Warentest" in

Germany), which is testing and rating consumer products; car buyers are likely checking J.D. Power's quality scores from the vendor's Initial Quality Study (IQS, problems after 3 months) and Vehicle Dependability Study (VDS, problems after 3 years); home buyers are worried about credit scores (FICO in the US, SCHUFA in Germany) and familiar with credit scoring agencies, such as Equifax; and finally, bond investors watch ratings of creditworthiness of corporate bonds ratings from specialists, like Moody's (Aa1) and Standard & Poor's (AA+).

As of 2019, for data, similar quality scoring solutions are missing. On one hand, this is a surprise because the importance of data quality seems to be unequivocally acknowledged, and therefore, scoring is recognized as a core data factory module (as illustrated in Fig. 2). On the other hand, data may be too complex with varying flavors across domains and industries to warrant an easy solution, which in turn presents a business opportunity.

The data science community confirms the old adage of "garbage in, garbage out" (GIGO): "Dirty data" is seen as the most common problem for workers in data science according to a survey with 16,000 responses on Kaggle (Kaggle 2017). With data analytics all insights are extracted from inside the data. Therefore, it is imperative to ensure that any raw data used has the information required for insights in it. One analogy is iron ore: For iron one would need rocks so rich with iron oxides that metallic iron can be extracted. Without iron oxide in it a rock would simply be a rock not iron ore.

Yet, despite the importance of data quality much work remains custom, hand-made and qualitative. The literature is using concepts, such as the "3 Vs" of volume, velocity and variety (McAfee and Brynjolfsson 2012); and more Vs are being added, like variability and value (e.g., Yin and Kaynak 2015). However, from an operational perspective, from an analytics application point of view, the Vs have remained conceptual and qualitative. The Vs may be useful for a first assessment, maybe for a pre-test, a first triage type data selection. However, in order to gauge outcomes in terms of performance, to estimate the likelihood of effects (x improves y), the size of effects (x improves y by a lot) and significance (improvements are real not random), the Vs have too little information in them. For example, consider traffic data for a routing app or parking data for a parking app: How fresh is the data? How frequently updated? Or consider time series data: How long and granular is it? Length of the overall observation period (10 years as opposed to 1 year), and for any period how dense is the data (one year's worth of data in monthly, daily, hourly intervals?). Our survey confirms the quality rating opportunity: While quality is preferred over quantity (How to spend the next US$1? Quality: 82%, quantity: 18%; n = 65), no obvious quality indicator is emerging (volume: 3%, freshness: 30%, format: 34%, source: 31%, license type: 2%, n = 64).

2.6 Governance

Many AI applications, such as predictive maintenance or autonomous driving, can require more data than what is available within a single department and company. Creating data pools across companies would be an advantage (IDSA 2019, Fig. 2.3, p. 15). For example, pooling all data of a particular machine type across all installations (horizontal pooling) would create a rich dataset for anomaly detection and its root-cause analysis.

Another use case is pooling data vertically, across the participants along an entire supply chain or channel system in order to better estimate arrival times or ensure proper end-to-end temperature treatment of shipments, for example. In both situations, horizontal and vertical pools, outcomes would be best if most participants were to contribute. However, so far, few companies have been willing to engage in this type of data sharing. On one hand, data is increasingly seen as a strategic advantage (the value aspect of "data is the new oil"), and therefore, held closely and protected. On the other hand, more sensor data will only increase data pooling benefits. What has been missing are exchange options with data governance mechanisms that strike a balance between the need to protect one's data and share it with others (Otto et al. 2016; IDSA 2018a, 2018b).

Such data governance solutions are emerging. An important example is the reference architecture model (RAM) of the International Data Spaces Association (IDSA 2019). IDSA is an association of industry participants, created to promote data governance architecture solutions based on research conducted by German Fraunhofer Institute with funding from the German government (Fraunhofer 2015).

Members include automakers like Volkswagen, suppliers like Bosch, and traditional information technology specialists like IBM.

The core element of IDSA RAM is a "connector." It ensures that data rights can be governed. Figure 3 illustrates the role of this element in the data flow between source (data provider) and sink (data consumer). With a connector any data package or product can be "wrapped up" in instructions and rules for use. Technically, it is a dedicated software component allowing participants to exchange, share and process data such that the data sovereignty of the data owner can be guaranteed.

Depending on the type of configuration, the connector's tamper-proof runtime can host a variety of system services including secure bidirectional communication, enforcement of content usage policies (e.g., expiration times and mandatory deletion of data), system monitoring, and logging of content transactions for clearing purposes.

As illustrated in Fig. 3, the functional range of a connector may be extended by (a) custom data apps, such as data visualization, provided in an app store and (b) a broker function to allow for product listings, such as a marketplace menu, and clearing services. A first connector implementation has been certified by IDSA for Deutsche Telekom's Data Intelligence Hub (Fraunhofer 2019).

3 Concluding Comments

Data and data analytics are seen as the next big business opportunity. Yet, today, data analytics is handmade and the overwhelming share of the time budget of a data analytics project is spent on refining data. In order to boost productivity and to prepare for even more IoT data while minimizing the risk of non-compliance with regulation, a "data factory" is required for data productization in an automated manner. Building on a foundation in the literature we have added bookends on data rights management and data governance in response to emerging data regulation. A data factory can be internal or external: It can be operated internally within the IT function (e.g., under a Chief Information Officer, CIO) or outside of it (e.g., under a Chief Marketing Officer, CMO), or it can be a separate, standalone business entirely. First standalone data factory service offerings

have already arrived with large enterprises, for example in Microsoft's Azure cloud and in the Telekom Data Intelligence Hub. Internal data factories may provide a way forward to extract value from data lakes and convert cost into business advantage by creating data products for internal operations and applications, such as anomaly detection, or into top-line growth with the sale of data products to third parties. Finally, combining a data factory with a data exchange may be an elegant way for large multi-divisional companies to quickly enable and promote a data-centric organization across functional or departmental silos. It's a classic: Top down push versus bottom up pull. Instead of having to define details upfront and top down, such as data product types and quality standards, and to enforce cooperation across silos, the market forces of an exchange would weed out lemon products and make data attributes and quality transparent.

References

Akerlof, G.A.: The market for 'lemons': quality uncertainty and the market mechanism. Q. J. Econo. **84**(3), 488–500 (1970)

American Marketing Association The 2018 AMA gold top 50 report (2018). https://www.ama.org/marketing-news/the-2018-ama-gold-top-50-report/

BMJV, Federal Ministry of Justice and Consumer Protection German Civil Code BGB (2013). http://www.gesetze-im-internet.de/englisch_bgb/index.html

Clark, K.B., Fujimoto, T.: Product Development Performance: Strategy, Organization, and Management in the World Auto Industry. In: Harvard Business School Press: Boston, MA (1991)

Clemons, E.K., Krcmar, H., Hermes, S., Choi, J.: American domination of the net: a preliminary ethnographic exploration of causes, economic implications for Europe, and future prospects. In: 52nd Hawaii International Conference on System Sciences (HICSS) (2019). https://doi.org/10.24251/hicss.2019.737

Crosby, L., Langdon, C.S.: Data as a product to be managed. Mark. News Am. Mark. Assoc. (2019). https://www.ama.org/marketing-news/data-is-a-product

Crosby, L, Langdon, C.S.: Technology personified. Mark. News Am. Mark. Assoc. 18–19 (2014)

Deutsche Telekom At a glance (2019). https://www.telekom.com/en/company/at-a-glance

Deutsche Telekom Creating value: Deutsche Telekom makes data available as a raw material. Press Release (2018). https://www.telekom.com/en/media/media-information/archive/deutsche-telekom-makes-data-available-as-a-raw-material-542866

Dewenter, R., Lueth, H.: Datenhandel und Plattformen. Gutachten. In: ABIDA – Assessing Big Data, German Federal Ministry of Education and Research (2018)

European Commission General Data Protection Regulation (2018). https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

Fraunhofer: Hannover Tradefair 2019, 'International Data Space'-Architecture implemented in first digital ecosystems. Fraunhofer Institute for Software and Systems Engineering, Press Release (2019). https://www.isst.fraunhofer.de/en/events/InternatData_Space-Architecture_implemented_in_first_digital_ecosystems.html

Fraunhofer: Fraunhofer initiative for secure data space launched. Fraunhofer Society for the Advancement of Applied Research, Press Release (2015). https://www.fraunhofer.de/en/press/research-news/2015/september/Fraunhofer-initiative-for-secure-data-space-launched.html

Gartner: Gartner top 10 strategic technology trends for 2019 (2018a). https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019/

Gartner: Gartner forecasts worldwide public cloud revenue to grow 17.3 percent in 2019 (2018b). https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019

Glassberg Sands, E.: How to build great data products. In: Harvard Business Review (2018). https://hbr.org/2018/10/how-to-build-great-data-products

IDSA Reference architecture model. International Data Spaces Association, Version 3.0 (2019). https://www.internationaldataspaces.org/info-package/

IDSA Sharing data while keeping data ownership, the potential of IDS for the data economy. International Data Spaces Association, White Paper (October) (2018a). https://www.internationaldataspaces.org/publications/sharing-data-while-keeping-data-ownership-the-potential-of-ids-for-the-data-economy/

IDSA Jointly paving the way for a data driven digitisation of European industry. International Data Spaces Association, White Paper Version 1.0 (October) (2018b). https://www.internationaldataspaces.org/publications/strategic-paper-for-europe-idsa/

Kaggle The state of data science & machine learning (2017). https://www.kaggle.com/surveys/2017

Krzanich, B.: Data is the new oil in the future of automated driving. Intel Newsroom (2016). https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving/#gs.6pqyxh

McAffee, A., Brynjolfsson, E.: Big data the management revolution. Harvard Bus. Rev. **90**(10), 60–68 (2012)

McKinsey Global Institute Notes from the AI frontier - insights from hundreds of use cases. McKinsey & Company (2018). https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

McKinsey & Company Monetizing car data. Advanced Industries Report (2016). https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/monetizing-car-data

Melendez, S.: A landmark Vermont law nudges over 120 data brokers out of the shadows. Fast Company (2019). https://www.fastcompany.com/90302036/over-120-data-brokers-inch-out-of-the-shadows-under-landmark-vermont-law

Microsoft Introduction to azure data factory (2018). https://docs.microsoft.com/en-us/azure/data-factory/introduction

Miller, R.: AWS and Microsoft reap most of the benefits of expanding cloud market. In: Techcrunch (2019). https://techcrunch.com/2019/02/01/aws-and-microsoft-reap-most-of-the-benefits-of-expanding-cloud-market/

Oregon State University Research data services: Data types & file formats (2019). https://guides.library.oregonstate.edu/research-data-services/data-management-types-formats

Otto, B., Juerjens, J., Schon, J., Auer, S., Menz, N., Wenzel, S., Cirullies, J.: Industrial data space - digital sovereignty over data. Fraunhofer Society for the Advancement of Applied Research (Working Paper) (2016). https://www.fraunhofer.de/content/dam/zv/en/fields-of-research/industrial-data-space/whitepaper-industrial-data-space-eng.pdf

Otto, B.: Organizing data governance: findings from the telecommunications industry and consequences for large service providers. Commun. AIS **29**(1), 45–66 (2011)

Palmer, M.: Data is the new oil. CMO News, ANA - Association of National Advertisers (2006). https://ana.blogs.com/maestros/2006/11/data_is_the_new.html

Pekka Pääkkönen, P., Pakkala, D.: Reference architecture and classification of technologies, products and services for big data systems. Big Data Res. **2**, 166–186 (2015)

Porter, M.E., Heppelmann, J.E.: How smart, connected products are transforming companies. Harvard Bus. Rev. (2015). https://hbr.org/2015/10/how-smart-connected-products-are-transforming-companies

Press, G.: Cleaning big data: Most time-consuming, least enjoyable data science task, Survey Says. Forbes (2016)

Schlueter Langdon, C.: Designing information systems capabilities to create business value: a theoretical conceptualization of the role of flexibility and integration. J. Data. Manage. **17**(3), 1–18 (2006)

Schlueter Langdon, C.: Information systems architecture styles and business interaction patterns: toward theoretic correspondence. J. Inf. Syst. E-Bus. **1**(3), 283–304 (2003a)

Schlueter Langdon, C.: The state of web services. IEEE Comput. **36**(7), 93–95 (2003b)

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V.: Critical analysis of big data challenges and analytical methods. J. Bus. Res. **70**, 263–286 (2017)

Sorge, P.: Germany backs european cloud project to avoid dependence on US technology. Wall Street J. (2019). https://www.wsj.com/articles/BT-CO-20190924-705704

Stieglitz, S., Mirbabayea, M., Rossa, B., Neubergerb, C.: Social media analytics – challenges in topic discovery, data collection, and data preparation. Int. J. Inf. Manage. **39**, 156–168 (2018)

Stern, D., Herbrich, R., Graepel, T.: Matchbox: large scale bayesian recommendations. In: Proceedings of the 18th International World Wide Web Conference (2009). https://www.microsoft.com/en-us/research/publication/matchbox-large-scale-bayesian-recommendations/?from=http%3A%2F%2Fresearch.microsoft.com%2Fpubs%2F79460%2Fwww09.pdf

USGS Data & file formats. United States Geological Survey (2019). https://www.usgs.gov/products/data-and-tools/data-management/data-file-formats

Vermont Office of the Attorney General Guidance on Vermont's Act 171 of 2018 Data Broker Regulation (2018). https://ago.vermont.gov/wp-content/uploads/2018/12/2018-12-11-VT-Data-Broker-Regulation-Guidance.pdf

Vollenweider, M.: Mind + Machine: A decision model for optimization and implementing analytics. John Wiley & Sons: Hoboken, NJ (2016)

Wall, M.: Tech trends 2019: 'The end of truth as we know it?' In: BBC (2019). https://www.bbc.com/news/business-46745742

Walter, S.M., Böhmann, T., Krcmar, H.: Industrialisierung der IT — Grundlagen, Merkmale und Ausprägungen eines Trends. HMD Praxis der Wirtschaftsinformatik **44**(4), 6–16 (2007). https://doi.org/10.1007/BF03340302

Womack, J., Jones, D., Roos, D.: The Machine that Changed the World: The Story of Lean Production. Free Press, Simon & Schuster, New York, NY (1990)

Yin, S., Kaynak, O.: Big data for modern industry: challenges and trends. Proc. IEEE **103**(2), 143–146 (2015)

Zuboff, S.: The Age of surveillance capitalism: the fight for a human future at the new frontier of power. New York, NY, Hachett Book Group (2019)

Zuckerberg, M.: The facts about facebook - We need your information for operation and security, but you control whether we use it for advertising. Wall Street J. (2019). https://www.wsj.com/articles/the-facts-about-facebook-11548374613