DOI:10.1145/3448247

Industry experiences on the data challenges of AI and the call for a data ecosystem for industrial enterprises.

BY CHRISTOPH GRÖGER

There Is No AI Without Data

ARTIFICIAL INTELLIGENCE (AI) has evolved from hype to reality over the past few years. Algorithmic advances in machine learning and deep learning, significant increases in computing power and storage, and huge amounts of data generated by digital transformation efforts make AI a game-changer across all industries.[8] AI has the potential to radically improve business processes with, for instance, real-time quality prediction in manufacturing, and to enable new business models, such as connected car services and self-optimizing machines. Traditional industries, such as manufacturing, machine building, and automotive, are facing a fundamental change: from the production of physical goods to the delivery of AI-enhanced processes and services as part of Industry 4.0.[25] This paper focuses on AI for industrial enterprises with a special emphasis on machine learning and data mining.

Despite the great potential of AI and the large investments in AI technologies undertaken by industrial enterprises, AI has not yet delivered on the promises in industry practice. The core business of industrial enterprises is not yet AI-enhanced. AI solutions instead constitute islands for isolated cases—such as the optimization of selected machines in the factory—with varying success. According to current industry surveys, data issues constitute the main reasons for the insufficient adoption of AI in industrial enterprises.[27,35]

In general, it is nothing new that data preparation and data quality are key for AI and data analytics, as there is no AI without data. This has been an issue since the early days of business intelligence (BI) and data warehousing.[3] However, the manifold data challenges of AI in industrial enterprises go far beyond detecting and repairing dirty data. This article profoundly investi-

» key insights

- **Despite AI's great potential, the business of industrial enterprises is not yet AI-enhanced. AI is done in an insular fashion, leading to a polyglot and heterogeneous enterprise data landscape that limits the comprehensive application of AI.**
- **Data challenges, such as data management, data democratization, and data governance, constitute the major obstacles to leveraging AI and go far beyond ensuring data quality, comprising diverse aspects such as metadata management, data architecture, and data ownership.**
- **The presented data ecosystem for industrial enterprises addresses these challenges and comprises data producers, data platforms, data consumers, and data roles for AI.**

IMAGE BY ALBERTO ANDREI ROSU, WITH ADDITIONAL IMAGERY FROM SHUTTERSTOCK.COM

gates these challenges and rests on our practical real-world experiences with the AI enablement of a large industrial enterprise—a globally active manufacturer. At this, we undertook systematic knowledge sharing and experience exchange with other companies from the industrial sector to present common issues for industrial enterprises beyond an individual case.

As a starting point, we characterize the current state of AI in industrial enterprises, called "insular AI," and present a practical example from manufacturing. AI is typically done in islands for use case-specific data provisioning and data engineering, leading to a heterogeneous and polyglot enterprise data landscape. This causes various data challenges that limit the comprehensive application of AI.

We particularly investigate challenges to data management, data democratization, and data governance which result from real-world AI projects. We illustrate them with practical examples and systematically elaborate on related aspects, such as metadata management, data architecture, and data ownership. To address the data challenges, we introduce the data ecosystem for industrial enterprises as an overall framework. We detail both IT-technical and organizational elements of the data ecosystem—for example, data platforms and data roles. Next, we assess how the data ecosystem addresses individual data challenges and paves the way from insular AI to industrialized AI. Then, we highlight the open issues we face in the course of our real-world realization of the data ecosystem and point out future research directions—for instance, the design of an enterprise data marketplace.

Current State of AI in Industrial Enterprises

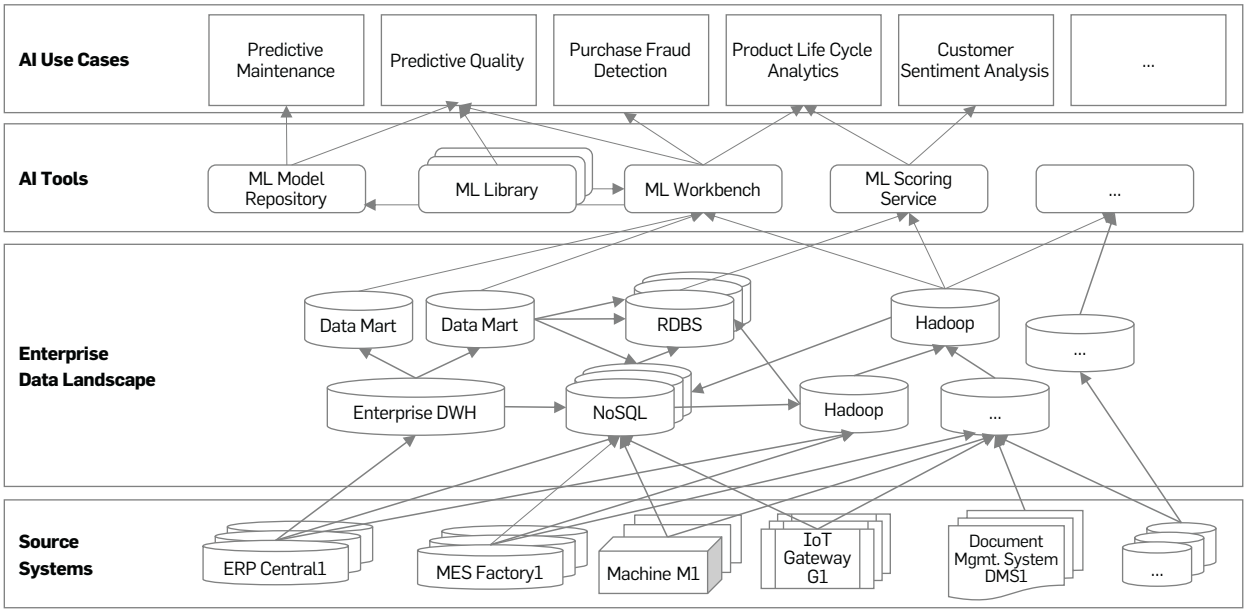
In the following, we define AI and data analytics as key terms and offer an overview of the business of industrial enterprises to concretize the scope of our work. On this basis, we characterize the current state of AI and illustrate it with a practical example.

Artificial intelligence and data analytics. Generally, AI constitutes a fuzzy term referring to the ability of a machine to perform cognitive functions.[10] Approaches to AI can be subdivided into deductive—that is, model-driven (such as expert systems)—or inductive—that is, data-driven.[10] In this paper, we focus on data-driven approaches, particularly machine learning and data mining,[17] as they have opened new fields of application for AI in the last years. Moreover, we use data analytics[4] as an umbrella term for all kinds of data-driven analysis, including BI and reporting.

Business of industrial enterprises. The business of industrial enterprises comprises the engineering and manufacturing of physical goods—for instance, heating systems or electrical drives. For this purpose, industrial enterprises typically operate a manufacturing network of various factories organized into business units. The IT landscape of industrial enterprises usually comprises different enterprise IT systems, ranging from enterprise resource planning (ERP) systems over product lifecycle management (PLM) systems to manufacturing execution systems (MES).[24] In Industry 4.0 and Internet of Things (IoT) applications, industrial enterprises push the digitalization of the industrial value chain.[22] The aim is to integrate data across the value chain and exploit it for competitive advantage. Hence, the AI enablement of processes and products is of strategic importance. To this end, industrial enterprises have, in recent years, built data lakes, introduced AI tools, and created data science teams.[15]

Current state: insular AI. Figure 1

Figure 1. Current state of AI in industrial enterprises: insular AI with heterogeneous enterprise data landscape.



DWH: Data Warehouse, ERP: Enterprise Ressource Planning, MES: Manufacturing Execution System, ML: Machine Learning, RDBS: Relational Database System

illustrates the current state of AI in industrial enterprises per the results of our investigations. Organizations have implemented a wide variety of AI use cases across the industrial value chain: from predictive maintenance for IoT-enabled products over predictive quality for manufacturing process optimization to product lifecycle analytics and customer sentiment analysis (see Gröger and Laudon, et al.[15,24] for details on these use cases). The use cases combine data from various source systems, such as ERP systems and MESs, and are typically implemented as isolated solutions for each individual case. That means, AI is performed in "islands" for use case-specific data provisioning and data engineering as well as for use case-specific AI tools and fit-for-purpose machine-learning algorithms. This is what we call "insular AI."

On one hand, insular AI fosters the flexibility and explorative nature of use-case implementations. On the other hand, it hinders reuse, standardization, efficiency, and enterprise-wide application of AI. The latter is what we call "industrialized AI." In the rest of this article, we focus on data-related issues of AI because the handling of data plays a central role on the path to industrialized AI. In fact, data handling accounts for around 60% to 80% of the entire AI use case implementation, according to our experiences.

Insular AI leads to a globally distributed, polyglot, and heterogeneous enterprise data landscape (see Figure 1). Structured and unstructured source data for AI use cases is extracted and stored in isolated raw data stores, called data lakes.[13] They are based on individual data storage technologies—for instance, different NoSQL systems, use case-specific data models, and dedicated source-data extracts. These data lakes coexist with the enterprise data warehouse,[23] which contains integrated and structured data from various ERP systems for reporting purposes. The many data-exchange processes in existence cause diverse data redundancies and potential data quality issues. Besides, the disparate data landscape significantly complicates the development of an integrated, enterprise-wide view of business objects—for example, products and processes—and thus hinders cross-process and cross-product AI use cases.

AI has not yet delivered on the promises in industry practice. The core business of industrial enterprises is not yet AI-enhanced.

Practical manufacturing example. To illustrate the shortcomings of insular AI and underline the need for an overall approach, we take an example from manufacturing. To predict the quality of a specific manufacturing process in a factory, a specialist project team of data scientists and data engineers first identifies relevant source systems, especially several local MESs in the factory as well as a central ERP system. The MESs provide sensor data on quality measurements and the ERP system provides master data. Together with various IT specialists, manufacturing experts, and data owners, the team inspects the data structures of the source systems and develops customized connectors for extracting source data and storing it in the local factory data lake in its raw format.

Data is cleansed, integrated, and pivoted based on a use case-specific data model and various case-specific data pipelines. As a general documentation of the business meaning of individual tables and columns is missing, this is done manually in the project's internal documents. The team then employs various machine-learning tools to generate an optimal prediction model. Over the course of several iterations, the data model and source-data extracts are adapted to enhance the data basis for machine learning. The final prediction model is then used in the MES on the factory shop floor by calling a machine-learning scoring service.

Overall, the resulting solution constitutes a targeted but isolated AI island with use case-specific data extracts, custom data models, tailored data pipelines, a dedicated factory data lake, and fit-for-purpose machine-learning tools. At this, the solution incorporates a large body of expert knowledge considering manufacturing-process know-how, ERP and MES IT system know-how, use case-specific data engineering, and data science know-how. Yet, missing data management guidelines (such as those for data modeling and metadata management), little transparency on source systems, and the variety of isolated data lakes all hinder reuse, efficiency, and enterprise-wide application of AI. That

is, the same type of use case gets implemented from scratch in different ways across different factories even though it refers to the same type of source systems, the same conceptual data entities, and the same type of manufacturing process. Thus, the same source data—for instance, master data—is extracted multiple times, creating a high load on business-critical source systems, such as ERP systems. Different data models are developed for the same conceptual data entities, such as 'machine' and 'product'. These heterogeneous data models and different data-storage technologies used in individual factory data lakes lead to heterogeneous data pipelines for pivoting the same type of source data, such as MES tables with sensor data. Besides, the business meaning of data and developed data models—that is, metadata—are documented multiple times in project-specific tools, such as data dictionaries or spreadsheets. All in all, this leads to an ocean of AI islands and a heterogeneous enterprise data landscape.

Consequently, to industrialize AI requires a systematic analysis of the underlying data challenges. On this basis, an overall solution integrating technical and organizational aspects can be designed to address the challenges.

Data Challenges of AI

Based on our practical investigations at the manufacturer, we identified manifold data challenges of AI and systematically clustered them. We aligned these challenges with other companies during systematic knowledge sharing to present common issues for industrial enterprises. Current literature[6,21] and industry surveys[27,35] on AI in industrial enterprises support our findings. Notably, this article goes significantly beyond these related works by analyzing both organizational and technical aspects of the data challenges and by providing detailed industry experiences on the individual challenges.

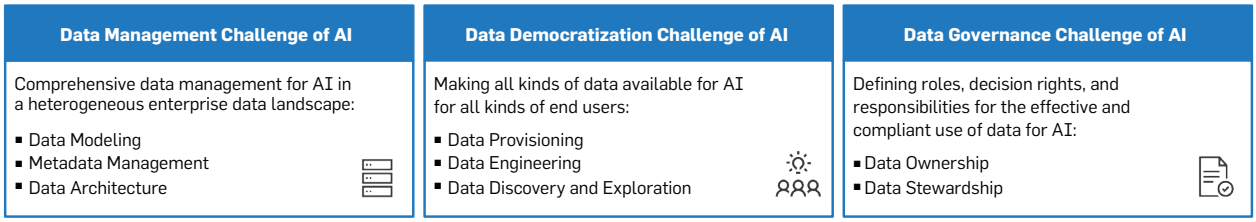
Generally, ensuring data quality for AI is important—for instance, by detecting and cleansing dirty data. Such data quality issues have already been addressed by a plurality of works and tools.[5,39] Beyond data quality, however, exist further critical data challenges—data management, data democratization, and data governance for AI (see Figure 2)—which we focus on in this article. We detail them with special emphasis on data-driven AI—that is, machine learning and data mining. In contrast to classical BI and reporting, machine learning and data mining impose extended data requirements.[6] They favor the use of not only aggregated, structured data but also of high volumes of both structured and unstructured data in its raw format—for example, for machine learning-based optical inspection.[40] This data also needs to be processed not only in periodic batches but also in near real time to provide timely results—for instance, to predict manufacturing quality in real time.[6] Consequently, AI poses new challenges to data management, data democratization, and data governance as detailed in the following.

According to current industry surveys, data issues constitute the main reasons for the insufficient adoption of AI in industrial enterprises.

Data management challenge of AI. Data management generally comprises all concepts and techniques to process, provision, and control data throughout its life cycle.[18] The data management challenge of AI lies in comprehensively managing data for AI in a heterogeneous and polyglot enterprise data landscape. According to our practical investigations, this particularly refers to data modeling, metadata management, and data architecture for AI.

No common data modeling approaches exist for how to structure and model data on a conceptual and logical level across the data landscape. Frequently, different data-modeling techniques, such as data vault[26] or dimensional modeling,[23] are used for the same kinds of data—for instance, manufacturing sensor data—in the data lakes. Sometimes, even the need for data modeling is neglected with reference to a flexible schema-on-read approach on top of raw data. This significantly complicates data integration, reuse of data, and developed data pipelines across different AI use cases. For instance, pivoting sensor data as input for machine learning is time-consuming and complex. Reusing corresponding data pipelines for different AI use cases significantly depends on common data-modeling techniques and common data models for manufacturing data, in this example.

Figure 2. Data challenges of AI and related aspects.



There is no overall metadata management to maintain metadata across the data landscape. Technical metadata, such as the names of columns and attributes, are mostly stored in the internal data dictionaries of individual storage systems and are not generally accessible. Hence, data lineage and impact analyses are hindered. For instance, in the case of changes in source systems, manually adapting the affected data pipelines across all data lakes without proper lineage metadata is tedious and costly. Moreover, business metadata on the meaning of data—for example, the meaning of KPIs—is often not systematically managed at all. Thus, missing metadata management significantly hampers data usage for AI.

No overarching data architecture structures the data landscape. Missing on one hand is an enterprise data architecture to orchestrate various isolated data lakes. For instance, there is no common zone model[37] across all data lakes, which complicates data integration and exchange. Moreover, the integration of the existing enterprise data warehouse containing valuable key performance indicators (KPIs) for AI use cases is unclear. On the other hand, also lacking is a systematic platform data architecture to design a data lake. Specifically, different data storage technologies are used to realize data lakes. For example, some data lakes are solely based on Hadoop storage technologies, such as HDFS[a] and Hive,[b] while others combine classical relational database systems and NoSQL systems. This leads to non-uniform data-lake architectures across the enterprise data landscape, resulting in high development and maintenance costs.

Data democratization challenge of AI. In general, data democratization refers to facilitating the use of data by everyone in an organization.[41] The data democratization challenge of AI lies in making all kinds of data available for AI for all kinds of end users across the entire enterprise. To this end, data provisioning and data engineering as well as data discovery and exploration all play central roles for AI. According to our investigations, these activities are mostly limited to small groups of expert users in practice and thus prevent data democratization for AI as explained in the following.

Data provisioning—that is, technically connecting new source systems to a data lake and extracting selected source data—typically requires dedicated IT projects. To that end, IT experts are concerned with defining technical interfaces and access rights for source systems and developing data extraction jobs in cooperation with source-system owners and data end users. Hence, the central IT department frequently becomes a bottleneck factor for data provisioning in practice. Moreover, there is a huge need for coordination between IT experts, source-system owners, and end users, which leads to time-consuming iterations. These factors significantly slow down and limit data provisioning and thus the use of new data sources for AI.

Data engineering—modeling, integrating, and cleansing of data—is typically done by highly skilled data scientists and data engineers. Due to incomplete metadata on source systems, data engineering requires specialist knowledge of individual source systems and their data structures—for example, on the technical data structures of ERP systems. In addition, mostly complex, script-based frameworks, such as with Python,[c] are used for data-engineering tasks requiring comprehensive programming know-how. These factors limit data engineering to small groups of expert users.

This also holds true for data discovery and exploration. Although self-service visualization tools are provided, discovery and exploration of data in data lakes is hampered. Comprehensive metadata on the business meaning and quality of data is missing, preventing easy data usage by non-expert users. For instance, a marketing specialist must identify and contact several different data engineers, who have prepared different kinds of market data, to understand the meaning and interrelations of the data. Besides, compliance approvals for data usage are typically based on specialist inspections of data, such as inspections by legal experts in the case of personal data. These low-automation processes also slow down the use of data for AI.

Data governance challenge of AI. Generally, data governance is about creating organizational structures to treat data as an enterprise asset.[1] The data governance challenge of AI refers to defining roles, decision rights, and responsibilities for the economically effective and compliant use of data for AI. According to our practical investigations, organizational structures for data are only rudimentarily implemented in industrial enterprises and mainly focus on master data and personal data. Particularly, structures for data ownership and data stewardship are missing, hampering the application of AI as follows.

There is no uniform data ownership organization across the heterogeneous data landscape. Especially, data own-

a http://hive.apache.org
b http://hadoop.apache.org
c http://www.python.org

ership for data extracted and stored in different data lakes is not defined in a common manner. For instance, in many cases, the owner of the data in the data lake remains the same as the data owner of the source system. That is, the integration of data from different source systems stored in the data lake requires approvals by different data owners. Hence, data is not treated as an enterprise asset owned by the company but rather as an asset of an individual business function—for example, the finance department as data owner of finance data. This leads to unclear responsibilities and an unbalanced distribution of risks and benefits when using data for AI.

For example, when manufacturing-process data from an MES is integrated with business-process data from an ERP system to enable predictive maintenance, the respective data owners—for instance, the manufacturing department and the finance department—must agree on and remain liable for a possibly noncompliant use of this data. However, the benefit of a successful use-case implementation, such as lower machine-maintenance costs, is attributed to the engineering department. In other cases, data ownership in the data lake is decoupled from data ownership in source systems to avoid this issue. Yet, this may lead to heterogeneous and overlapping data ownership structures, such as when data ownership is organized by business function in source systems and by business unit in the data lake. These organizational boundaries significantly hinder the comprehensive use of data for AI.

There is no overall data stewardship organization to establish common data policies, standards, and procedures. Existing data stewardship structures in industrial enterprises mainly focus on various kinds of master data to define—for example, common data quality criteria for master data on customers. Data stewardship for further categories of data is not systematically organized. For example, there are various data models as well as data quality criteria on manufacturing data across different factories and manufacturing processes. Thus, common enterprise-wide policies for manufacturing data are lacking. This significantly increases the efforts and complexity of data engineering for AI use cases.

Call for a Data Ecosystem for Industrial Enterprises

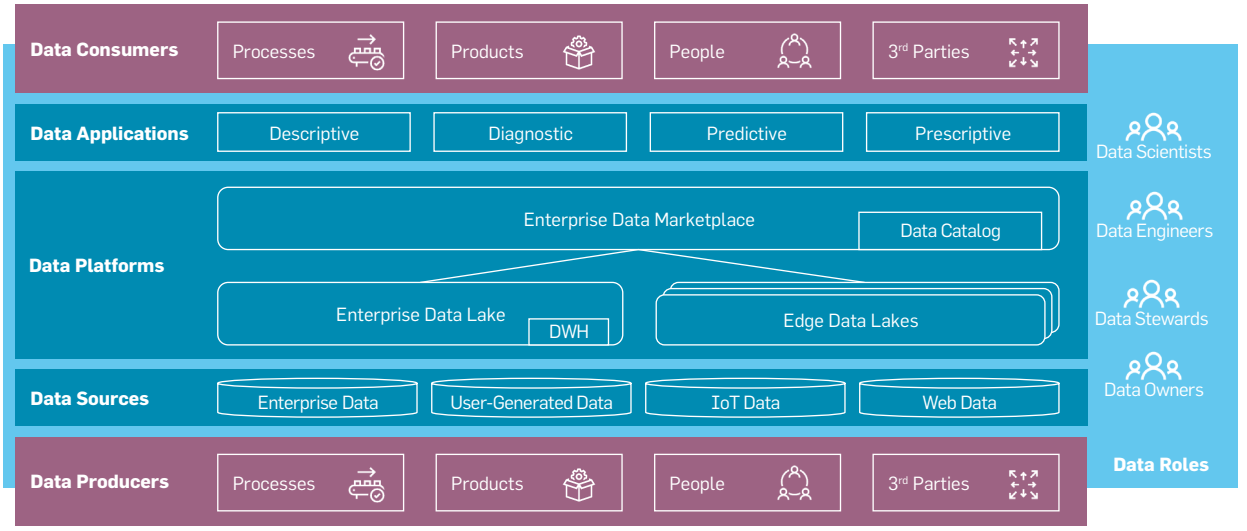
In light of the above data challenges, we see the need for a holistic framework that covers both technical and organizational aspects to address the data challenges of AI. To this end, we adopt the framework of a data ecosystem. Generally, a data ecosystem represents a socio-technical, self-organizing, loosely coupled system for the sharing of data.[31] A data ecosystem's typical elements are data producers, data consumers, and data platforms.[31] However, data ecosystem research is still in its early stages and mainly focused on the sharing of open government data.[33] Therefore, we call for a data ecosystem specifically tailored to industrial enterprises.

Based on our practical experiences with the AI enablement of the manufacturer and knowledge exchange with further industrial companies, we derived core data ecosystem elements for industrial enterprises (see Figure 3). They are described in the following:

Data producers and data consumers. Data producers and consumers represent resources or actors generating or consuming data. We generally differentiate four kinds of data producers in an industrial enterprise: Processes refer to all kinds of industrial processes and resources across the value chain—for instance, engineering processes.[24] Products refer to manufactured goods, such as electrical drives or household appliances. People comprise all kinds of human actors, including customers and employees. Third parties comprise actors and resources outside the organizational scope of the enterprise—for example, suppliers.

Data sources. Data sources relate to the technical kind and the sources of data generated by data producers. We distinguish between four kinds of data

Figure 3. Core elements of a data ecosystem for industrial enterprises.



sources in an industrial enterprise: Enterprise data refers to all data generated by enterprise IT systems across the industrial value chain, such as PLM and ERP systems.[24] User-generated data refers to data directly generated by human actors, such as social media postings or documents. IoT data refers to all data generated by IoT devices, such as manufacturing machine data or sensor data.[6] Web data refers to all data from the Web, except user-generated data—for instance, linked open data or payment data.

Data platforms. Data platforms represent the technical foundation for data processing from all kinds of data sources to make data available for various data applications. The data ecosystem is based on three kinds of data platforms: the enterprise data lake, edge data lakes, and the enterprise data marketplace.

The enterprise data lake constitutes a logically central, enterprise-wide data lake. It combines the original data lake approach[29] with the data warehouse concept.[23] That means, it combines the data lake-like storage and processing of all kinds of raw data with the data warehouse-like analysis of aggregated data. Batch and stream data processing are supported to enable all kinds of analyses on all kinds of data. The enterprise data lake is based on comprehensive guidelines for data modeling and metadata management and enables enterprise-wide reuse of data and data pipelines.

Edge data lakes represent decentralized raw data stores that complement the enterprise data lake. Edge data lakes focus on the realization of data applications based on local data, with little enterprise-wide reuse. They are particularly suited for data processing in globally distributed factories, with selected factories operating their own edge data lake. A typical AI use case for edge data lakes is to predict time-series data produced by a specific manufacturing machine in a single factory of the enterprise.

The enterprise data marketplace constitutes the central pivot point of the data ecosystem. It represents a metadata-based self-service platform that connects data producers with data consumers. The goal is to match supply and demand for data within the enterprise. However, research on data marketplaces is at an early stage and there are only initial concepts focusing on external enterprise marketplaces for data.[36,38] Hence, we work out essential characteristics of an internal enterprise data marketplace fitting the data ecosystem.

Based on our practical experiences with the AI enablement of the manufacturer and knowledge exchange with further industrial companies, we derived core data ecosystem elements for industrial enterprises.

In contrast to the enterprise data lake and edge data lakes, the enterprise data marketplace does not store the actual data. Rather, it is based on a data catalog[37] representing a metadata-based inventory of data. That is, data is represented by metadata and a reference to the actual data. For instance, the data catalog item, "Quality Data for Product P71" might comprise metadata on the related product and a reference to a set of sensor data stored in the enterprise data lake. Data catalog items not only refer to data in the data lakes but also to data in source systems, such as ERP and PLM systems. Besides, metadata from application programming interfaces (APIs) that expose data are also fused in the data catalog. Hence, the marketplace in combination with the data catalog provides a metadata-based overview of all data in the enterprise.

Regarding services provided by the marketplace, it addresses both data consumers and data producers in a self-service manner. Data consumer services comprise things like self-service data discovery and self-service data preparation. Data producer services include, for instance, self-service data curation to define metadata on datasets as well as self-services for API-based data publishing. Marketplace services on the whole address the entire data lifecycle: data acquisitioning and cataloging, publishing and lineage tracking, and data preparation and exploration.

Data applications. Data applications refer to all kinds of applications that use data provided by the data platforms. We differentiate descriptive, diagnostic, predictive, and prescriptive data applications.[15] That is, data applications comprise the entire range of data analytics techniques, from reporting to machine learning. Data applications realize defined use cases, such as process performance prediction in manufacturing, for defined data consumers—for instance, a process engineer.

Data roles. Data roles comprise

organizational roles related to data. These roles are relevant across all layers of the data ecosystem. We focus on key roles that are of central importance for AI and data analytics in industrial enterprises—namely data owners, data stewards, data engineers, and data scientists.

Data owners[1] have the overall responsibility for certain kinds of data—for instance, all data on a certain product. They are assigned to the business, not IT, and are responsible for the quality, security, and compliance of this data from a business point of view. It is particularly important to define a uniform and transparent data ownership organization across the enterprise data lake and the edge data lakes and to decouple these structures from data ownership in source systems. For instance, all data on a specific product stored in the enterprise data lake should be owned by the respective business unit, to facilitate cross-process use of data.

Data stewards[1] manage data on behalf of data owners. They are responsible for realizing necessary policies and procedures from both business and technical points of view. To reduce the complexity and efforts of data engineering for AI, an overall data stewardship organization is needed, establishing common quality criteria and reference data models for all kinds of data. For instance, manufacturing data can be structured according to the IEC 62264 reference model[20] to ease data integration across different factories of the enterprise.

Data engineers and data scientists are key roles within the context of AI projects but there is no widely accepted definition—yet.[28] Generally, data engineers develop data pipelines to provide the data basis for further analyses by integrating and cleansing data. Building on this foundation, data scientists focus on actual data analysis by feature engineering and applying various data analytics techniques—for instance, different machine-learning algorithms—to derive insights from data.

From Insular AI to Industrialized AI: Addressing Challenges and Future Directions

We are currently realizing the data ecosystem on an enterprise-scale at the manufacturer to evolve from insular AI to industrialized AI. Generally, the data ecosystem paves the way to industrialized AI by addressing the data challenges. To assess this, we analyze individual data challenges with respect to data ecosystem elements (see Table). We highlight open issues we are facing during the course of our real-world realization of the data ecosystem and point out future research directions. Further details on the realization of selected elements of the data ecosystem can be found in our most recent works.[12–16]

We see a major need for future research regarding functional capabilities and realization technologies for an enterprise data marketplace.

Addressing the data management challenge. With respect to the data management challenge, the data ecosystem is based on a comprehensive set of data platforms, namely the enterprise data lake, edge data lakes, and the enterprise data marketplace. These platforms define an enterprise data architecture for AI and data analytics, specifically addressing the aspect of data architecture. For this purpose, the enterprise data lake incorporates the enterprise data warehouse, avoiding two separate enterprise-wide data platforms and corresponding data redundancies. It is based on a unified set of data modeling guidelines and reference data models implemented by data stewards to address the aspect of data modeling. For instance, enterprise data from ERP systems is modeled using data vault modeling to enable rapid integration with sensor data from IoT devices as described in our recent work.[14] This enables the enterprise-wide reuse of data and data pipelines for all kinds of AI use cases across products, processes, and factories. Additionally, edge data lakes provide flexibility for use-case exploration and prototyping with only minimal guidelines, but they are restricted to local data, particularly in single factories.

The design of the platform data architecture of the enterprise data lake itself is challenging, as it must serve a huge variety of data applications, from descriptive reporting to predictive and prescriptive machine-learning applications. Particularly, defining a suitable composition of data storage and processing technologies is an open issue. According to our practical experiences, the enterprise data lake favors a polyglot approach to provide fit-for-purpose technologies for different data applications. To this end, we combine

relational database systems, NoSQL systems, and real-time event hubs following the lambda architecture paradigm as discussed in our recent work.[15]

Identifying suitable architecture patterns for different kinds of data applications on top of this polyglot platform constitutes a valuable future research direction for standardizing the implementation of AI use cases. In addition, organizing all data in the enterprise data lake requires an overarching structure beyond conceptual data modeling. We see data lake zones[37] as a promising approach necessitating substantial future research as discussed in our recent work.[12]

The aspect of metadata management is addressed by the data catalog as part of the enterprise data marketplace. The data catalog focuses on the acquisition, storage, and provisioning of all kinds of metadata—technical, business, and operational—across all data lakes and source systems. In this way, it enables overarching lineage analyses and data quality assessments as essential parts of AI use cases—for example, to evaluate the provenance of a dataset in the enterprise data lake. Data catalogs represent a relatively new kind of data management tool and mainly focus on the management of metadata from batch storage systems—such as relational database systems as detailed in our recent work.[13] Open issues particularly refer to the integrated management of metadata from batch and streaming systems, such as Apache Kafka, to realize holistic metadata management in the data ecosystem.

Addressing the data democratization challenge. All aspects of the data democratization challenge—namely data provisioning, data engineering, and data discovery and exploration—refer to self-service and metadata management. They are addressed by the enterprise data marketplace based on the data catalog. The data catalog provides comprehensive metadata management across all data lakes and source systems of the data ecosystem. Thus, it significantly facilitates data engineering as well as data discovery and exploration for all kinds of end users by providing technical and business information on data and its sources as discussed in our recent work.[16] For instance, the business meaning of calculated KPIs in the enterprise data lake can be investigated, and corresponding source systems can be looked up easily in the data catalog by non-expert users.

The enterprise data marketplace also provides self-service capabilities across the entire data lifecycle for all kinds of data producers and data consumers. For instance, a process engineer in manufacturing provisions sensor data of a new machine in the enterprise data lake himself by executing a self-service workflow in the data marketplace.

Neither established tools nor sound concepts for internal enterprise data marketplaces exist, hence we are realizing the marketplace as an individual software development project. To this end, there are various realization options—for instance, using semantic technologies for modeling metadata and services.[7] Thus, we see a major need for future research regarding functional capabilities and realization technologies for an enterprise data marketplace.

Addressing the data governance challenge. In view of the data governance challenge, the data ecosystem defines a set of key roles related to data—namely data owners, data stewards, data engineers, and data scientists. Thus, both aspects—data ownership and data stewardship—are addressed. An overarching data ownership organization across source systems and data lakes facilitates the compliant and prompt provisioning of source data for AI use cases because approvals and responsibilities for the use of data are clearly defined. Moreover, a data stewardship organization for all kinds of data significantly enhances data quality and reduces data engineering efforts by establishing reference data models and data quality criteria. At this, the data catalog supports data governance by providing KPIs for data owners and data stewards, such as the number of sources of truth for specific data sets.

A major open issue refers to the implementation of these roles within existing organizational structures. Generally, there are various data governance frameworks and maturity models in literature and practice.[1,2,9,18,19,30,32,34] However, they only provide high-level guidance on how to approach data governance—for example, what topics to address and what roles to define. Concrete guidelines covering how to implement data governance, considering context factors such as industry and corporate culture, are lacking—for instance, deciding when data ownership is to be organized by business unit

Addressing data challenges by the data ecosystem and resulting future research directions.

Data Challenges of AI	Aspects	Data Ecosystem Approach	Future Research Directions
Data Management Challenge of AI	Data Modeling	Unified data modeling concepts and reference data models in the enterprise data lake	Overall data organization in enterprise data lake—for instance, using data lake zones
	Metadata Management	Data catalog for metadata management	Integrated management of metadata from batch and streaming systems
	Data Architecture	Architecture consisting of enterprise data lake, edge data lakes, and enterprise data marketplace	Polyglot platform data architecture of enterprise data lake, including architecture patterns
Data Democratization Challenge of AI	Data Provisioning	Self-service and metadata management provided by enterprise data marketplace and data catalog	Framework of capabilities and realization technologies for an enterprise data marketplace
	Data Engineering		
	Data Discovery and Exploration		
Data Governance Challenge of AI	Data Ownership	Key roles for data owners, data stewards, data engineers, and data scientists	Implementation guidelines for data roles considering context factors—for example, corporate culture
	Data Stewardship		

or by business process.[1] Thus, we see a need for future research concerning context-based implementation guidelines for data roles.

Conclusion

Data challenges constitute the major obstacle to leveraging AI in industrial enterprises. According to our investigations of real-world industry practices, AI is currently undertaken in an insular fashion, leading to a polyglot and heterogeneous enterprise data landscape. This presents considerable challenges for systematic data management, comprehensive data democratization, and overall data governance and prevents the widespread use of AI in industrial enterprises.

To address these issues, we presented the data ecosystem for industrial enterprises as a guiding framework and overall architecture. Our assessment of the data challenges against the data ecosystem elements underlines that all data challenges are addressed—paving the way from insular AI to industrialized AI. The socio-technical character of the data ecosystem allows organizations to address both the technical aspects of the data management challenge and the organizational aspects of the data governance challenge—with defined data roles and data platforms. Furthermore, the loosely coupled and self-organizing nature of the data ecosystem with self-reliant data producers and data consumers addresses the data democratization challenge—for instance, with comprehensive self-service and metadata management provided by the enterprise data marketplace. At this, the data ecosystem is valid not only for AI but also for any kind of data analytics, as it addresses all types of data sources and all types of data applications in industrial environments. It is to be noted that the data ecosystem elements were derived from our practical findings and generalized for industrial enterprises. We encourage additional work to further refine and validate these elements.

We are currently realizing the data ecosystem at the manufacturer on an enterprise-scale and are facing various issues that indicate the need for further research. In particular, the design of an enterprise data marketplace as a novel type of data platform constitutes a valuable direction of future work.

Acknowledgments

The author would like to thank Jens Bockholt and Dieter Neumann for their continuous support of this work. Moreover, big thanks go to Arnold Lutsch and Eva Hoos for their valuable comments.

References

1. Abraham, R., Schneider, J., and Brocke, J.v. Data governance: A conceptual framework, structured review, and research agenda. *Intern. J. of Information Management 49* (2019), 424–438.
2. Ballard, C., Compert, C., Jesionowski, T., Milman, I., Plants, B., Rosen, B., and Smith, H. Information governance principles and practices for a big data landscape. *IBM* (2014).
3. Ballou, D.P. and Tayi, G.K. Enhancing data quality in data warehouse environments. *Communications of the ACM 42*, 1 (1999), 73–78.
4. Cao, L. Data science: A comprehensive overview. *ACM Computing Surveys 50*, 3 (2017), 1–42.
5. Chu, X., Ilyas, I.F., Krishnan, S., and Wang, J. Data cleaning: Overview and emerging challenges. In *Proceedings of the Intern. Conf. on Management of Data (SIGMOD)*, ACM, New York (2016), 2201–2206.
6. Cui, Y., Kara, S., and Chan, K.C. Manufacturing big data ecosystem: A systematic literature review. *Robotics and Computer Integrated Manufacturing 62* (2020) Article 101861.
7. Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., and Bartolucci, A. The advantages of an ontology-based data management approach: Openness, interoperability and data quality. *Scientometrics 108*, 1 (2016), 441–455.
8. Davenport, T.H. and Ronanki, R. Artificial intelligence for the real world. *Harvard Business Review 96*, 1 (2018), 108–116.
9. The DGI data governance framework. The Data Governance Institute (2020).
10. Everitt, T. and Hutter, M. Universal artificial intelligence. Practical agents and fundamental challenges. *Foundations of Trusted Autonomy*. H. Abbass, J. Scholz, and D. Reid, eds. Springer. (2018) 15–46.
11. Gessert, F., Wingerath, W., Friedrich, S., and Ritter, N. NoSQL database systems: A survey and decision guidance. *Computer Science—Research and Development 32*, 3–4 (2016), 353–365.
12. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. A zone reference model for enterprise-grade data lake management. In *Proceedings of the IEEE Enterprise Distributed Object Computing Conf. (EDOC)*, IEEE, Piscataway, New Jersey (2020), 57–66.
13. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. Leveraging the data lake: Current state and challenges. In *Proceedings of the Intern. Conf. on Big Data Analytics and Knowledge Discovery (DaWaK)*, Springer, Cham, (2019), 179–188.
14. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. Modeling data lakes with data vault: Practical experiences, assessment, and lessons learned. In *Proceedings of the Intern. Conf. on Conceptual Modeling (ER)*, Springer, Cham, (2019), 63–77.
15. Gröger, C. Building an Industry 4.0 analytics platform. *Datenbank-Spektrum 18*, 1 (2018), 5–14.
16. Gröger, C. and Hoos, E. Ganzheitliches metadatenmanagement im data lake: Anforderungen, IT-werkzeuge und herausforderungen in der praxis. In *Proceedings Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW)*, Gesellschaft für Informatik, Bonn, (2019), 435–452.
17. Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Amsterdam, (2012).
18. Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., and Smith, E. *DAMA-DMBOK: Data Management Body of Knowledge*. Technics Publications, New Jersey, (2017).
19. Holistic data governance: A framework for competitive advantage. Informatica (2017).
20. IEC 62264-2:2015. Enterprise-control system integration—Part 2: Objects and attributes for enterprise-control system integration. International Organization for Standardization (2015).
21. Ismail, A., Truong, H.-L., and Kastner, W. Manufacturing process data analysis pipelines: A requirements analysis and survey. *Journal of Big Data 6*, 1 (2019), 1–26.
22. Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., and Eschert, T. Industrial Internet of Things and cyber manufacturing systems. *Industrial Internet of Things*. S. Jeschke, C. Brecher, H. Song, and D. Rawat, eds. Springer, (2017), 3–19.
23. Kimball, R. and Ross, M. *The Data Warehouse Toolkit. The Definitive Guide to Dimensional Modeling*. Wiley, Indianapolis, (2013).
24. Laudon, K.C. and Laudon, J.P. *Management Information Systems. Managing the Digital Firm*. Pearson Education, Harlow, (2018).
25. Lee, J., Davari, H., Singh, J., and Pandhare, V. Industrial artificial intelligence for Industry 4.0-based manufacturing systems. *Manufacturing Letters 18*, (2018), 20–23.
26. Linstedt, D. and Olschimke, M. *Building a Scalable Data Warehouse with Data Vault 2.0*. Morgan Kaufmann, Waltham, (2016).
27. Loucks, J., Davenport, T.H., and Schatsky, D. State of AI in the enterprise, 2nd edition. Deloitte, (2018).
28. Lyon, L. and Mattern, E. Education for real-world data science roles (part 2): A translational approach to curriculum development. *Intern. J. of Digital Curation 11*, 2 (2016), 13–26.
29. Mathis, C. Data lakes. *Datenbank-Spektrum 17*, 3 (2017), 289–293.
30. Morabito, V. *Big Data and Analytics*. Springer, Cham, (2015).
31. Oliveira, M.I.S., Fatima Barros Lima, G.d., and Loscio, B.F. Investigations into data ecosystems: A systematic mapping study. *Knowledge and Information Systems 61*, 2 (2019), 589–630.
32. Plotkin, D. *Data Stewardship*. Morgan Kaufmann. (2014).
33. Reggi, L. and Dawes, S. Open government data ecosystems: Linking transparency for innovation with transparency for participation and accountability. In *Proceedings of the Intern. Conf. on Electronic Government (EGOV)*, Springer, Cham, (2016), 74–86.
34. The SAS data governance framework: A blueprint for success. SAS (2018).
35. Schaeffer, E., Wahrendorff, M., Narsalay, R.M., Gupta, A., and Hobräck, O. Turning possibility into productivity. *Accenture* (2018).
36. Schomm, F., Stahl, F., and Vossen, G. Marketplaces for data: An initial survey. *ACM SIGMOD Record 42*, 1 (2013), 15–26.
37. Sharma, B. *Architecting Data Lakes*. O'Reilly, Sebastopol, CA, (2018).
38. Smith, G., Ofe, H.A., and Sandberg, J. Digital service innovation from open data: Exploring the value proposition of an open data marketplace. In *Proceedings of the Hawaii Intern. Conf. on System Sciences (HICSS)*, IEEE, Piscataway, New Jersey, (2016), 1277–1286.
39. Taleb, I., Serhani, M.A., and Dssouli, R. Big data quality: A survey. In *Proceedings of the IEEE Intern. Congress on Big Data*, IEEE, Piscataway, New Jersey, (2018), 166–173.
40. Yang, Y., Pan, L., Ma, J., Yang, R., Zhu, Y., Yang, Y., and Zang, L. A high-performance deep-learning algorithm for the automated optical inspection of laser welding. *J. of Applied Sciences 10*, 3 (2020), 1–11.
41. Zeng, J. and Glaister, K.W. Value creation from big data: Looking inside the black box. *Strategic Organization 16*, 2 (2018), 105–140.

Christoph Gröger (christoph.groeger@de.bosch.com) is enterprise architect for data analytics at Bosch and a senior technical professional in Bosch's global data strategy team in Stuttgart, Germany.

Copyright held by author(s)/owner(s).
Publication rights licensed to ACM.

Watch the author discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/no-ai-without-data