There Is No AI Without Data

Industry Experiences on the Data Challenges of AI and the Call for a Data Ecosystem for Industrial Enterprises

Christoph Gröger
Robert Bosch GmbH
70469 Stuttgart, Germany
{firstname.lastname}@de.bosch.com

ABSTRACT

Artificial intelligence (AI) constitutes a game changer across all business sectors. This holds particularly true for industrial enterprises due to the large amounts of data generated across the industrial value chain. However, AI has not delivered on the promises in industry practice, yet. The core business of industrial enterprises is not yet AI-enhanced. In fact, data issues constitute the main reasons for the insufficient adoption of AI. This paper addresses these issues and rests on our practical experiences on the AI enablement of a large industrial enterprise. As a starting point, we characterize the current state of AI in industrial enterprises, which we call "insular AI". This leads to various data challenges limiting the comprehensive application of AI. We particularly investigate challenges on data management, data democratization and data governance resulting from real-world AI projects. We illustrate these challenges with practical examples and detail related aspects, e.g., metadata management, data architecture and data ownership. To address the challenges, we present the data ecosystem for industrial enterprises. It constitutes a framework of data producers, data platforms, data consumers and data roles for AI and data analytics in industrial environments. We assess how the data ecosystem addresses the individual data challenges and highlight open issues we are facing in course of the enterprise-scale realization of the data ecosystem. Particularly, the design of an enterprise data marketplace as pivotal point of the data ecosystem is a valuable direction of future work.

KEYWORDS

Artificial Intelligence, Data Management, Data Democratization, Data Governance, Data Ecosystem, Industry Experience

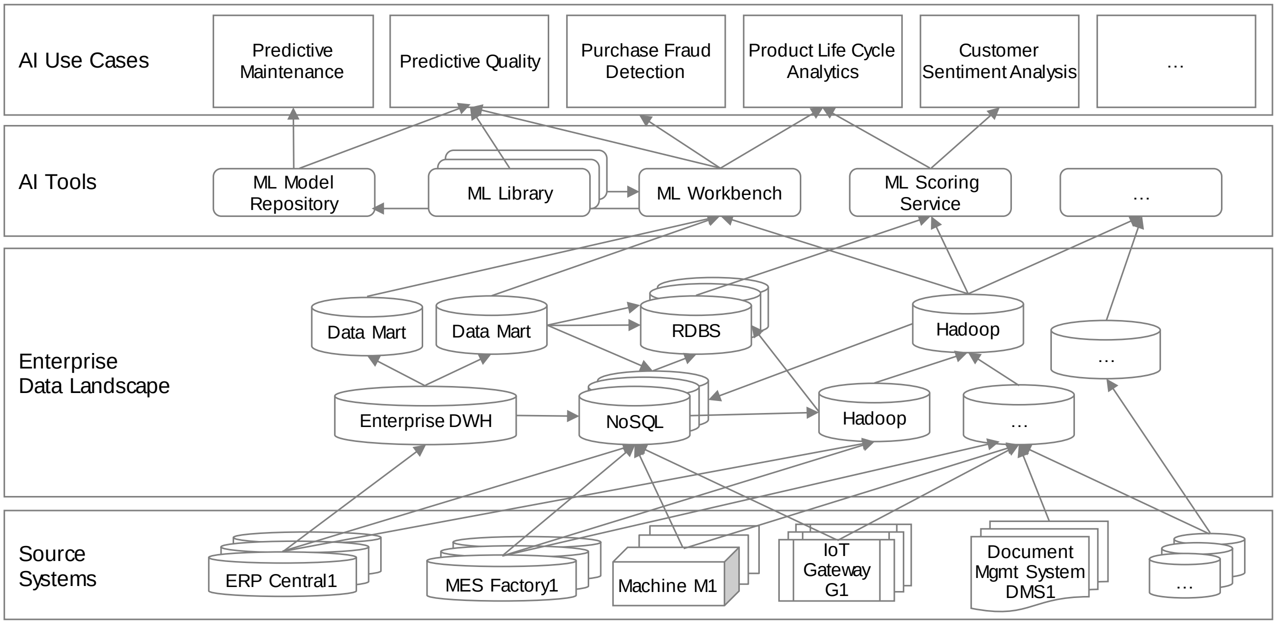
1 Introduction

In the last years, artificial intelligence (AI) has evolved from hype to reality. Algorithmic advances in machine learning and deep learning, significant increases in computing power and storage as well as huge amounts of data generated by the digital transformation make AI a game changer across all industries [8]. AI has the potential to radically improve business processes, e.g., by real-time quality prediction in manufacturing, and to enable new business models, e.g., connected car services and self-optimizing machines. In particular, traditional industries, such as manufacturing, machine building and automotive, are facing a fundamental change: from physical goods production to the delivery of AI-enhanced processes and services in course of industry 4.0 [25]. Thus, this paper focuses on AI for industrial enterprises with special emphasis on machine learning and data mining.

Despite the great potential of AI and the large investments industrial enterprises have undertaken in AI technologies, AI has not delivered on the promises in industry practice, yet. The core business of industrial enterprises is not yet AI-enhanced. AI solutions constitute rather islands for isolated cases, e.g., 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 and data warehousing [3]. However, the manifold data challenges of AI in industrial enterprises go far beyond detecting and repairing dirty data. The paper at hand profoundly investigates these challenges and rests on our practical real-world experiences on the AI enablement of a large industrial enterprise, a globally active manufacturer. At this, we did 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 (see Section 2). This causes various data challenges limiting the comprehensive application of AI. We particularly investigate challenges on data management, data democratization as well as data governance resulting 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 (see Section 3). 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, e.g., data platforms and data roles (see Section 4). Next, we assess how the data ecosystem addresses the individual data challenges and thereby paves the way from insular AI to industrialized AI. At this, we highlight open issues we



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

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

are facing in course of the real-world realization of the data ecosystem and point out future research directions, e.g., the design of an enterprise data marketplace (see Section 5). Finally, we conclude in Section 6.

2 Current State of AI in Industrial Enterprises

In the following, we define AI and data analytics as key terms and give 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.

2.1 Artificial Intelligence and Data Analytics

Generally, *AI* constitutes a fuzzy term referring to the ability of a machine to perform cognitive functions [10]. AI approaches can be subdivided into *deductive, model-driven* approaches, e.g., expert systems, as well as *inductive, data-driven* approaches [10]. In this paper, we focus on data-driven approaches, particularly *machine learning and data mining* [17], as they have opened up 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 approaches including business intelligence and reporting.

2.2 Business of Industrial Enterprises

The business of industrial enterprises comprises the engineering and manufacturing of physical goods, e.g., heating systems or electrical drives. For this purpose, industrial enterprises typically operate a manufacturing network of various factories organized in business units. The IT landscape of industrial enterprises is usually comprised of different enterprise IT systems ranging from enterprise resource planning (ERP) systems over product life cycle management (PLM) systems to manufacturing execution systems (MES) [24]. In course of industry 4.0 and the internet of things (IoT), industrial enterprises push the digitalization of the industrial value chain [22]. The aim is to integrate data across the value chain and exploit them for competitive advantages. Hence, the AI enablement of processes and products is of strategic importance. To this end, in the last years, data lakes were built up, AI tools were introduced and data scientist teams were established in industrial enterprises [15].

2.3 Current State: Insular AI

As a result of our investigations, the current state of AI in industrial enterprises is illustrated in Fig. 1. A wide variety of *AI use cases* has been implemented across the industrial value chain: from *predictive maintenance* for IoT-enabled products over *predictive quality* for manufacturing process optimization to *product life cycle analytics* and *customer sentiment analysis* (see [15, 24] for details on these use cases). The use cases combine data from various different *source systems*, e.g., *ERP* systems and *MES*, and are typically implemented as isolated solutions for each individual case.

That means, AI is done 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 the one hand, insular AI fosters the flexibility and the 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 the paper, we focus on data-related issues of AI because the handling of data plays a central role on the way to industrialized AI. In fact, data handling accounts for around sixty to eighty percent of the entire implementation effort of an AI use case according to our experiences.

Insular AI leads to a *globally distributed, polyglot and heterogeneous enterprise data landscape* (see Fig. 1). Structured and unstructured source data for AI use cases are extracted and stored in isolated raw data stores, i.e. *data lakes* [13]. They are based on individual data storage technologies, e.g., different NoSQL systems [11], 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. At this, a lot of data exchange processes exist causing diverse data redundancies and potential data quality issues. Besides, the disparate data landscape significantly complicates the development of an integrated enterprise-wide view on business objects, e.g., products and processes, and thus hinders cross-process and cross-product AI use cases.

2.4 Practical Example in Manufacturing

To illustrate the shortcomings of insular AI and underline the need for an overall approach, we take an example from manufacturing. For predicting 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 MES in the factory as well as a central ERP system. The MES 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 them in the local factory data lake in their raw format. Data are then 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 project-internal documents. Then, the team employs various different machine learning tools to stepwise generate an optimal prediction model. In course of several iterations, the data model and the 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 systems know-how as well as use-case-specific data engineering and data science know-how. Yet, missing data management guidelines, e.g., for data modelling and metadata management, little transparency on source systems as well as a variety of isolated data lakes 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 although 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, e.g., master data, are extracted multiple times creating 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, e.g., MES tables with sensor data. Besides, business meaning of data and developed data models, i.e., metadata, are documented multiple times in project-specific tools, e.g., 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, a systematic analysis of the underlying data challenges is necessary (see Section 3). On this basis, an overall solution approach can be designed that integrates IT-technical and organizational aspects to address the challenges (see Section 4).

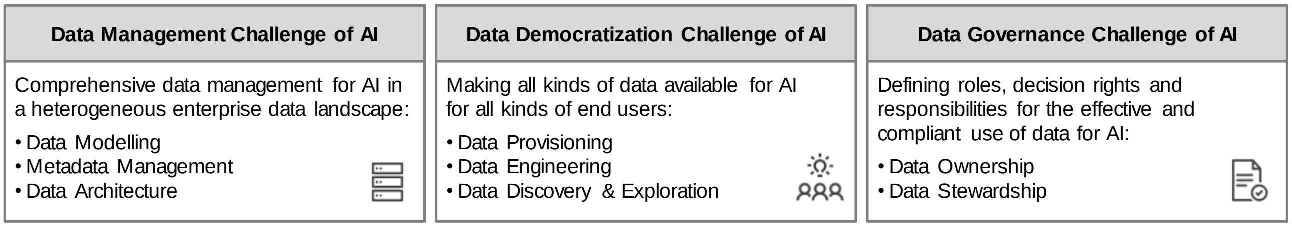
3 Data Challenges of AI

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

Generally, ensuring data quality for AI is important, e.g., by detecting and repairing dirty data. Such data quality issues have already been addressed by a plurality of works and tools [5, 39]. However, there are further critical data challenges beyond data quality, which we focus on in this paper. They refer to *data management*, *data democratization* and *data governance for AI* (see Fig. 2). We detail them below with special emphasis on data-driven AI, i.e., machine learning and data mining. In contrast to classical business intelligence and reporting, machine learning and data mining impose extended data requirements [6]. They favor the use of not only aggregated, structured data but of high volumes of both structured and unstructured data in their raw format, e.g., for machine-learning-based optical inspection [40]. Besides, these data need to be processed not only in periodic batches but also near real-time to provide timely results, e.g., for real-time manufacturing quality prediction [6]. Consequently, AI poses new challenges to data management, data democratization and data governance as detailed in the following.

3.1 Data Management Challenge of AI

Data management generally comprises all concepts and techniques to process, provision and control data throughout its lifecycle [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 modelling*, *metadata management* and *data architecture* for AI as explained in the following.



Data Management Challenge of AI

Comprehensive data management for AI in a heterogeneous enterprise data landscape:

- Data Modelling
- Metadata Management
- Data Architecture

Data Democratization Challenge of AI

Making all kinds of data available for AI for all kinds of end users:

- Data Provisioning
- Data Engineering
- Data Discovery & Exploration

Data Governance Challenge of AI

Defining roles, decision rights and responsibilities for the effective and compliant use of data for AI:

- Data Ownership
- Data Stewardship

Fig. 2. Data challenges of AI and related aspects

There are *no common data modelling approaches* on how to structure and model data on a conceptual and logical level across the data landscape. Frequently, *different data modelling techniques*, e.g., data vault [26] or dimensional modelling [23], are used for the same kinds of data, e.g., 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 and reuse of data and developed data pipelines across different AI use cases. For instance, pivoting sensor data as input for a machine learning is time-consuming and complex. Reusing corresponding data pipelines for different AI use cases significantly depends on common data modelling techniques and common data models for manufacturing data, in this example.

There is *no overall metadata management* to maintain metadata across the data landscape. *Technical metadata*, e.g., names of columns and attributes, are mostly stored in system-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 is tedious and costly without proper lineage metadata. Moreover, *business metadata* on the meaning of data, e.g., the meaning of KPIs, are often not systematically managed, at all. Thus, missing metadata management significantly hampers data usage for AI as detailed in Section 3.2.

There is *no overarching data architecture* structuring the data landscape. On the one hand, an *enterprise data architecture* to orchestrate the various isolated data lakes is missing. For instance, there is no common zone model [37] across all data lakes complicating data integration and exchange. Moreover, the integration of the existing enterprise data warehouse containing valuable KPIs for AI use cases is unclear. On the other hand, a systematic *platform data architecture* to design a data lake itself is lacking. In particular, different data storage technologies are used to realize data lakes. For example, some data lakes are solely based on Hadoop[1] storage technologies, e.g., HDFS[1] and Hive[2], 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.

3.2 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*, *data engineering* as well as *data discovery and exploration* for AI play a central role. 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, i.e., technically connecting new source systems to a data lake and extracting selected source data, typically requires dedicated IT projects. At these, 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 the end users of data. 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 the end users leading 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, i.e., modelling, integrating and cleansing of data, is typically done by highly skilled data scientists and data engineers. Due to incomplete metadata on source systems (see Section 3.1), data engineering requires specialist knowledge on individual source systems and their data structures, e.g., on technical data structures of ERP systems. In addition, mostly complex script-based frameworks, e.g., with Python[3], are used for data engineering tasks requiring comprehensive programming know-how. These factors limit data engineering to small groups of expert users.

For *data discovery and exploration*, this also holds true. Although self-service visualization tools are provided, discovery and exploration of data in data lakes are hampered. Comprehensive metadata on the business meaning and the quality of data are missing and thus prevent easy data usage by non-expert users. For instance, a marketing specialist has to identify and contact several different data engineers, who prepared different kinds of market data, in order to understand the meaning and the interrelations of these data. Besides, compliance approvals for data usage are typically based on specialist inspections of data, e.g., inspections by

[1] http://hadoop.apache.org
[2] http://hive.apache.org
[3] http://www.python.org

legal experts in the case of personal data. These low-automation processes slow down the use of data for AI, as well.

3.3 Data Governance Challenge of AI

Generally, data governance is about 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 rudimentary implemented in industrial enterprises and mainly focus on master data and personal-related data. Particularly, structures for *data ownership* and *data stewardship* are missing and hamper the application of AI as follows.

There is *no uniform data ownership organization* across the heterogeneous data landscape. Especially, data ownership for data extracted and stored in the 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 that are stored in the data lake requires approvals by different data owners. Hence, data are not treated as an enterprise asset owned by the company but rather as an asset of an individual business function, e.g., 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 are integrated with business process data from an ERP system to enable predictive maintenance, the respective data owners, e.g., the manufacturing department and the finance department, have to agree and remain liable for a possibly incompliant use of these data. However, the benefit of a successful use case implementation, e.g., reduced machine maintenance costs, are 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, e.g., 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 policies, standards and procedures for data. Existing data stewardship structures in industrial enterprises mainly focus on various kinds of master data, e.g., to define 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 different 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.

4 Call for a Data Ecosystem for Industrial Enterprises

In the light of the above data challenges, we see the need for a holistic framework covering both IT-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 and loosely coupled system for the sharing of data* [31]. Typical elements of a data ecosystem are data producers, data consumers and data platforms [31]. However, data ecosystem research is still at an early stage and mainly focuses 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 on the AI enablement of the manufacturer and knowledge exchange with further industrial companies, we derived core data ecosystem elements for industrial enterprises (see Fig. 3). They are described in the following.

4.1 Data Producers and Data Consumers

Data producers and *data 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 employed resources across the value chain, e.g., engineering processes [24]. *Products* refer to manufactured goods, e.g., electrical drives or household appliances. *People* comprise all kinds of human actors, e.g., customers and employees. *Third parties* comprise actors and resources outside the organizational scope of the enterprise, e.g., suppliers.

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

4.3 Data Platforms

Data platforms represent the technical foundation for data processing from all kinds of data sources in order to make data available for various data applications. The data ecosystem is based on three kinds of data platforms, namely 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. At this, batch and stream data processing are supported in order to enable all kinds of analyses on all kinds of data. The enterprise data lake is based on comprehensive guidelines for data modelling and metadata management and enables enterprise-wide reuse of data and data pipelines.

Edge data lakes represent decentral raw data stores that are used as complements to the enterprise data lake. Edge data lakes

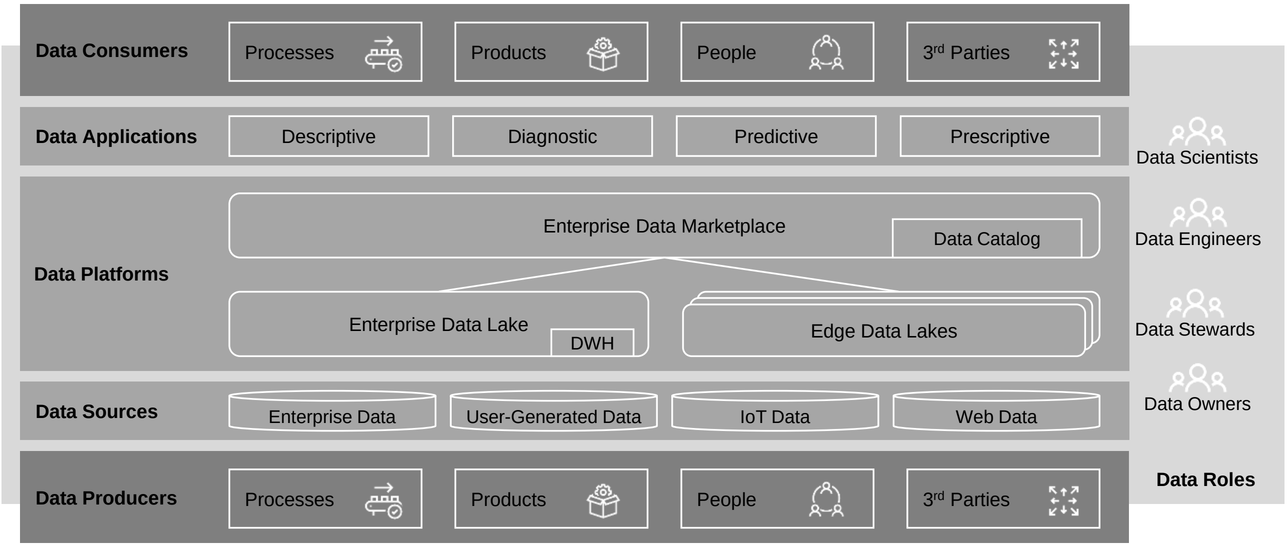


Fig. 3. Core elements of a data ecosystem for industrial enterprises

focus on the realization of data applications based on local data with little enterprise-wide reuse. Edge data lakes 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 the prediction of time series data produced by a specific manufacturing machine in a single factory of the enterprise.

The *enterprise data marketplace* constitutes the central pivotal point of the data ecosystem. It represents a metadata-based self-service platform that connects data producers and 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 company-external marketplaces for data [36, 38]. Hence, we work out essential characteristics of an enterprise-internal data marketplace fitting the data ecosystem.

In contrast to the enterprise data lake and the 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 are represented by metadata and a reference to the actual data. For instance, the data catalog item "Quality Data for Product P71" could be comprised of 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, e.g., ERP and PLM systems. Besides, metadata from application programming interfaces (APIs) that expose data are fused in the data catalog, as well. Hence, the marketplace in combination with the data catalog provides a metadata-based overview on 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, e.g., 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. As a whole, the marketplace services address the entire data lifecycle, from data acquisitioning and cataloging over publishing and lineage tracking to preparation and exploration of data.

4.4 Data Applications

Data applications refer to all kinds of applications that make use of 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, e.g., process performance prediction in manufacturing, for defined data consumers, e.g., a process engineer.

4.5 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, e.g., all data on a certain product. They are assigned to the business, not the IT, and are responsible for the quality, security and compliance of these 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 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, e.g., the electrical drives business unit, not by source system data owners in manufacturing or engineering, to facilitate cross-process use of data.

Data Challenges of AI	Aspects	Data Ecosystem Approach	Future Research Directions
Data Management Challenge of AI	Data Modelling	Unified data modelling concepts and reference data models in the enterprise data lake	Overall data organization in enterprise data lake, e.g., 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 & 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, e.g., corporate culture
	Data Stewardship		

Table 1: Addressing of data challenges by the data ecosystem and resulting future research directions

Data stewards [1] manage data on behalf of data owners. They are responsible for realizing necessary policies and procedures from a business and from a technical point 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 in the context of AI projects but there is no widely accepted definition, yet [28]. Generally, *data engineers* are responsible for developing data pipelines in order to provide the data basis for further analyses by integrating and cleansing of data. Building on this foundation, *data scientists* focus on the actual analysis of data by feature engineering and applying various data analytics techniques, e.g., different machine learning algorithms, to derive insights from data.

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

We are currently realizing the data ecosystem on an enterprise-scale at the manufacturer in order 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 the individual data challenges with respect to the data ecosystem elements (see Table 1). At this, we highlight open issues we are facing in 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].

5.1 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 (see Section 4.3). These platforms define an enterprise data architecture for AI and data analytics specifically addressing the aspect *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 modelling guidelines and reference data models implemented by data stewards in order to address the aspect *data modelling*. For instance, enterprise data from ERP systems are modelled using data vault to enable rapid integration with sensor data from IoT devices as described in our recent work [14]. This enables the enterprise-wide (re)use of data and data pipelines for all kinds of AI use cases across products, processes and factories. In addition, edge data lakes provide flexibility for use case exploration and prototyping with only minimal guidelines but 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 has to 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 in order to standardize the implementation of AI use cases. In addition, organizing all data in the enterprise data lake requires an overarching structure beyond conceptual data modelling. We see data lake zones [37] as a promising approach necessitating substantial future research as discussed in our recent work [12].

The aspect *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 part of AI use cases, e.g., 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, e.g., Apache Kafka, to realize a holistic metadata management in the data ecosystem.

5.2 Addressing the Data Democratization Challenge

All aspects of the data democratization challenge – namely *data provisioning*, *data engineering* as well as *data discovery and exploration* – refer to self-service and metadata management. They are addressed by the enterprise data marketplace based on the data catalog. As described Section 5.1, 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. In addition, the enterprise data marketplace provides self-services across the entire data lifecycle for all kinds of data producers and data consumers as detailed in Section 4.3. 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.

As there are neither established tools nor sound concepts for enterprise-internal data marketplaces, yet (see Section 4.3), we are realizing the marketplace as an individual software development project. At this, there are various realization options, e.g., using semantic technologies for modelling metadata and services [7]. Thus, we see a major need for future research regarding functional capabilities and realization technologies for an enterprise data marketplace.

5.3 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. As detailed in Section 4.5, an overall data ownership organization across source systems and data lakes facilitates the compliant and promptly 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, e.g., 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, e.g., what topics to address and what roles to define. Concrete guidelines how to implement data governance considering context factors, such as industry and corporate culture, are lacking, e.g., to decide when data ownership is to be organized by business unit or by business process [1]. Thus, we see a need for future research concerning context-based implementation guidelines for data roles.

6 Conclusion

Data challenges constitute the major obstacle to leverage AI in industrial enterprises. According to our investigations of real-world industry practice, AI is currently done in an insular fashion leading to a polyglot and heterogeneous enterprise data landscape. This makes systematic data management, comprehensive data democratization and an overall data governance considerably challenging and prevents the wide-spread 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 to address both the IT-technical aspects of the data management challenge and the organizational aspects of the data governance challenge, e.g., 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, e.g., 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 all kinds 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 novel type of data platform constitutes a valuable direction of future work.

ACKNOWLEDGEMENTS

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. (2019): Data governance: a conceptual framework, structured review, and research agenda. *International Journal of Information Management.* 49, 424–438.

[2] Ballard, C., Compert, C., Jesionowski, T., Milman, I., Plants, B., Rosen, B. and Smith, H. (2014): Information governance principles and practices for a big data landscape. IBM.

[3] Ballou, D.P. and Tayi, G.K. (1999): Enhancing data quality in data warehouse environments. *Communications of the ACM.* 42, 1, 73–78.

[4] Cao, L. (2017): Data science: a comprehensive overview. *ACM Computing Surveys.* 50, 3, 1–42.

[5] Chu, X., Ilyas, I.F., Krishnan, S. and Wang, J. (2016): Data cleaning: overview and emerging challenges. *Proceedings of the International Conference on Management of Data (SIGMOD)*, ACM, New York, 2201–2206.

[6] Cui, Y., Kara, S. and Chan, K.C. (2020): Manufacturing big data ecosystem: a systematic literature review. *Robotics and Computer Integrated Manufacturing.* 62, Article 101861.

[7] Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A. and Bartolucci, A. (2016): The advantages of an ontology-based data management approach: openness, interoperability and data quality. *Scientometrics.* 108, 1, 441–455.

[8] Davenport, T.H. and Ronanki, R. (2018): Artificial Intelligence for the real world. *Harvard Business Review.* 96, 1, 108–116.

[9] DGI (2020): The DGI data governance framework. The Data Governance Institute.

[10] Everitt, T. and Hutter, M. (2018): Universal artificial intelligence. Practical agents and fundamental challenges. *Foundations of trusted autonomy.* H. Abbass, J. Scholz, and D. Reid, eds. Springer. 15–46.

[11] Gessert, F., Wingerath, W., Friedrich, S. and Ritter, N. (2016): NoSQL database systems: a survey and decision guidance. *Computer Science - Research and Development.* 32, 3–4, 353–365.

[12] Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2020): A zone reference model for enterprise-grade data lake management. *Proceedings of the IEEE Enterprise Distributed Object Computing Conference (EDOC)*, IEEE, Piscataway, 57–66.

[13] Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2019): Leveraging the data lake: current state and challenges. *Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*, Springer, Cham, 179–188.

[14] Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2019): Modeling data lakes with data vault: practical experiences, assessment, and lessons learned. *Proceedings of the International Conference on Conceptual Modeling (ER)*, Springer, Cham, 63–77.

[15] Gröger, C. (2018): Building an industry 4.0 analytics platform. *Datenbank-Spektrum.* 18, 1, 5–14.

[16] Gröger, C. and Hoos, E. (2019): Ganzheitliches Metadatenmanagement im Data Lake: Anforderungen, IT-Werkzeuge und Herausforderungen in der Praxis. *Proceedings Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW)*, Gesellschaft für Informatik, Bonn, 435–452.

[17] Han, J., Kamber, M. and Pei, J. (2012): *Data mining: Concepts and techniques.* Morgan Kaufmann, Amsterdam.

[18] Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E. and Smith, E. (2017): *DAMA-DMBOK: Data management body of knowledge.* Technics Publications, New Jersey.

[19] Informatica (2017): Holistic data governance: a framework for competitive advantage.

[20] International Organization for Standardization (2015): IEC 62264-2:2015. Enterprise-control system integration - Part 2: Objects and attributes for enterprise-control system integration.

[21] Ismail, A., Truong, H.-L. and Kastner, W. (2019): Manufacturing process data analysis pipelines: a requirements analysis and survey. *Journal of Big Data.* 6, 1, 1–26.

[22] Jeschke, S., Brecher, C., Meisen, T., Özdemir, D. and Eschert, T. (2017): Industrial internet of things and cyber manufacturing systems. *Industrial internet of things.* S. Jeschke, C. Brecher, H. Song, and D. Rawat, eds. Springer. 3–19.

[23] Kimball, R. and Ross, M. (2013): *The data warehouse toolkit. The definitive guide to dimensional modeling.* Wiley, Indianapolis.

[24] Laudon, K.C. and Laudon, J.P. (2018): *Management information systems. Managing the digital firm.* Pearson Education, Harlow.

[25] Lee, J., Davari, H., Singh, J. and Pandhare, V. (2018): Industrial artificial intelligence for industry 4.0-based manufacturing systems. *Manufacturing Letters.* 18, 20–23.

[26] Linstedt, D. and Olschimke, M. (2016): *Building a scalable data warehouse with data vault 2.0.* Morgan Kaufmann, Waltham.

[27] Loucks, J., Davenport, T.H. and Schatsky, D. (2018): State of AI in the Enterprise, 2nd Edition. Deloitte.

[28] Lyon, L. and Mattern, E. (2016): Education for real-world data science roles (part 2): a translational approach to curriculum development. *International Journal of Digital Curation.* 11, 2, 13–26.

[29] Mathis, C. (2017): Data Lakes. *Datenbank-Spektrum.* 17, 3, 289–293.

[30] Morabito, V. (2015): *Big data and analytics.* Springer, Cham.

[31] Oliveira, M.I.S., Fatima Barros Lima, G. d. and Loscio, B.F. (2019): Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems.* 61, 2, 589–630.

[32] Plotkin, D. (2014): *Data stewardship.* Morgan Kaufmann, Waltham.

[33] Reggi, L. and Dawes, S. (2016): Open government data ecosystems: linking transparency for innovation with transparency for participation and accountability. *Proceedings of the International Conference on Electronic Government (EGOV)*, Springer, Cham, 74–86.

[34] SAS (2018): The SAS data governance framework: a blueprint for success. SAS Institute.

[35] Schaeffer, E., Wahrendorff, M., Narsalay, R.M., Gupta, A. and Hobräck, O. (2018): Turning possibility into productivity. Accenture.

[36] Schomm, F., Stahl, F. and Vossen, G. (2013): Marketplaces for data: an initial survey. *ACM SIGMOD Record.* 42, 1, 15–26.

[37] Sharma, B. (2018): *Architecting data lakes.* O'Reilly, Sebastopol.

[38] Smith, G., Ofe, H.A. and Sandberg, J. (2016): Digital service innovation from open data: exploring the value proposition of an open data marketplace. *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, IEEE, Piscataway, 1277–1286.

[39] Taleb, I., Serhani, M.A. and Dssouli, R. (2018): Big data quality: a survey. *Proceedings of the IEEE International Congress on Big Data (BigData Congress)*, IEEE, Piscataway, 166–173.

[40] Yang, Y., Pan, L., Ma, J., Yang, R., Zhu, Y., Yang, Y. and Zang (2020): A high-performance deep learning algorithm for the automated optical inspection of laser welding. *Journal of Applied Sciences.* 10, 3, 1–11.

[41] Zeng, J. and Glaister, K.W. (2018): Value creation from big data: Looking inside the black box. *Strategic Organization.* 16, 2, 105–140.