ICS 35.240.01

VDI/VDE-RICHTLINIEN

Oktober 2022
October 2022

VEREIN DEUTSCHER INGENIEURE

VERBAND DER ELEKTROTECHNIK ELEKTRONIK INFORMATIONSTECHNIK

Implementierung und Betrieb von Big-Data-Anwendungen in der produzierenden Industrie

Datenbewirtschaftung

Implementation and operation of big data applications in the manufacturing industry

Data management

VDI/VDE 3714

Blatt 3 / Part 3

Ausgabe deutsch/englisch
Issue German/English

Die deutsche Version dieser Richtlinie ist verbindlich.

The German version of this standard shall be taken as authoritative. No guarantee can be given with respect to the English translation.

Inhalt Seite

Contents Page

VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA)
Fachbereich Digitale Transformation

VDI-Handbuch Informationstechnik, Band 1: Angewandte Informationstechnik
VDI/VDE-Handbuch Automatisierungstechnik

Vorbemerkung

Der Inhalt dieser Richtlinie ist entstanden unter Beachtung der Vorgaben und Empfehlungen der Richtlinie VDI 1000.

Allen, die ehrenamtlich an der Erarbeitung dieser Richtlinie mitgewirkt haben, sei gedankt.

Eine Liste der aktuell verfügbaren und in Bearbeitung befindlichen Blätter dieser Richtlinienreihe sowie gegebenenfalls zusätzliche Informationen sind im Internet abrufbar unter www.vdi.de/3714.

Einleitung

Der Fachausschuss „Big Data“ der VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik hat sich mit der Erstellung dieser Richtlinie der Aufgabe angenommen, den ökonomischen und ökologischen Nutzen von Big Data aufzuzeigen, den Wissenstransfer über verschiedene Industrien und Branchen hinweg zu verbessern und die Implementierung und den Betrieb von Big-Data-Anwendungen in der produzierenden Industrie voranzutreiben und zu vereinheitlichen.

Die Richtlinienreihe soll eine Orientierung über erforderliche Maßnahmen zur Big-Data-Analyse geben und aufzeigen, welche Methoden für eine zielführende Arbeit geeignet sind und welche Einschränkungen und Hindernisse bestehen. Praktikern und Praktikerinnen sollen Hinweise gegeben werden, welche Methoden und Betrachtungen für den Erfolg eines Big-Data-Projekts hinsichtlich Einsatz und nachhaltigen Betrieb notwendig sind.

Die Richtlinienreihe VDI/VDE 3714 umfasst die Blätter:

Blatt 1 Durchführung von Big-Data-Projekten
Blatt 2 Datenqualität
Blatt 3 Datenbewirtschaftung
Blatt 4 Analyseverfahrensklassen
Blatt 5 Modellierungsverfahren
Blatt 6 Validierung von Modellen
Blatt 7 Online-Anwendung von datengetriebenen Modellen

Die Richtlinienreihe VDI/VDE 3714 ist im Fachausschuss 7.24 „Big Data“ des Fachbereichs 7 „Anwen-

Preliminary note

The content of this standard has been developed in strict accordance with the requirements and recommendations of the standard VDI 1000.

We wish to express our gratitude to all honorary contributors to this standard.

A catalogue of all available parts of this series of standards and those in preparation as well as further information, if applicable, can be accessed on the internet at www.vdi.de/3714.

Introduction

The “Big Data” Technical Committee of the VDI/VDE Society Measurement and Automatic Control has taken on the task of drawing up this standard to demonstrate the economic and ecological benefits of big data, to improve the transfer of knowledge across different industries and sectors, and to promote and standardize the implementation and operation of big data applications in the manufacturing industry.

This series of standards is intended to provide orientation on required measures for big data analysis and to show which methods are suitable for target-oriented work or which limitations and obstacles exist. Practitioners should be given advice on which methods and considerations are necessary for the success of a big data project regarding its use and sustainable operation.

The series of standards VDI/VDE 3714 comprises the parts:

Part 1 Implementation of big data projects
Part 2 Data quality
Part 3 Data management
Part 4 Analysis process classes
Part 5 Modelling procedures
Part 6 Validation of models
Part 7 Online application of data-driven models

The series of standards VDI/VDE 3714 is published in the Technical Committee 7.24 “Big Data” of the

dungsfelder der Automation“ der VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA) entstanden. Damit stellen die Produktion sowie die Mess- und Automatisierungstechnik die Schwerpunkte dar. In den Produktionsprozessen werden beispielsweise für Steuerungs- und Regelungsaufgaben oder für die Qualitätssicherung große Datenmengen erhoben, die mittels Datenanalyse für weitere Prozess- und Geschäftsverbesserungen genutzt werden können. Die Richtlinienreihe gibt eine generelle Orientierung sowie Hinweise auf potenzielle Schwierigkeiten und Hürden bei der Durchführung von Big-Data-Anwendungen – von der Entwicklung über die Inbetriebnahme bis zum nachhaltigen Betrieb.

Technical Division 7 “Application fields of automation” of the VDI/VDE Measurement and Automatic Control (GMA). Thus, production as well as measurement and automation technology represent the focal points. In production processes, for example, large amounts of data are collected for control and regulation tasks or for quality assurance, which can be used for further process and business improvements by means of data analysis. This series of standards provides a general orientation as well as indications of potential difficulties and hurdles in the implementation of big data applications, from development through commissioning to sustainable operation.

Benachbart zu Big Data finden sich Themen wie das Internet der Dinge (IoT), die Vernetzung von Geräten (Smart Devices) oder die zunehmende „Rechnerallgegenwart“ (Ubiquitous Computing) sowie Begriffe, wie Business Intelligence, Data Analytics, Advanced Analytics, Data Mining, Smart Data und Data-Warehouse-Systeme, die generell die Nutzung von Daten adressieren.

Adjacent to big data are topics such as the Internet of things (IoT), the networking of devices (smart devices), or the increasing “computer omnipresence” (ubiquitous computing), as well as terms such as business intelligence, data analytics, advanced analytics, data mining, smart data, and data warehouse systems that generally address the use of data.

Die Richtlinienreihe geht von einer generellen Verfügbarkeit aller benötigten Daten aus. Bezüglich der Datenmenge, ihrer Struktur und Integrität wird keine Annahme getroffen. Zur Diskussion und Charakterisierung der Daten helfen die sogenannten „fünf Vs“, die die einzelnen Dimensionen von Big Data bezeichnen. Die Daten werden durch Umfang (*Volume*), Unterschiedlichkeit (*Variety*) und ihre zeitliche Taktung (*Velocity*) charakterisiert. Insbesondere bei industriellen Anwendungen sind die Qualität der Daten (*Validity*) und der unternehmerische Mehrwert (*Value*) relevant.

The series of standards assumes a general availability of all required data. No assumption is made regarding the amount of data, its structure and integrity. For the discussion and characterization of data, the so-called “five Vs”, which denote the individual dimensions of big data, are helpful. The data is characterized by *volume*, *variety*, and *velocity*. The quality of the data (*validity*) and the added business value (*value*) are particularly relevant for industrial applications.

Auf weitere grundsätzliche technische Regeln sei hier hingewiesen, insbesondere im Umfeld von Industrie 4.0:

- VDI 2222 Blatt 1
- VDI/VDE 3517
- VDI/VDE 4000 Blatt 1
- VDI 4010
- DIN EN ISO 9000
- DIN EN ISO 9001
- DIN EN ISO 9004
- ISO 13053

Further fundamental technical rules should be pointed out here, especially in the environment of Industry 4.0:

- VDI 2222 Part 1
- VDI/VDE 3517
- VDI/VDE 4000 Part 1
- VDI 4010
- DIN EN ISO 9000
- DIN EN ISO 9001
- DIN EN ISO 9004
- ISO 13053

Um vorhandene Daten aus der Produktion einer Analyse zugänglich zu machen, ist es in den meisten Fällen notwendig, die Daten aus verschiedenen Datenquellen zusammenzuführen und in einer geeigneten Weise so zu transformieren, dass eine Datenanalyse möglich wird.

In order to make existing data from production accessible for analysis, it is necessary in most cases to combine the data from different data sources and to transform it in a suitable way so that data analysis becomes possible.

Die hierfür benötigten technischen Prozesse werden als ETL-Prozesse bezeichnet, wobei ETL für *Extract*, *Transform* und *Load* steht. Datenbewirtschaftungsarchitekturen moderner Big-Data-Anwendungen sehen häufig ELT-Prozesse (*Extract*, *Load*, *Transform*) vor. Diese unterscheiden sich von ETL-Prozessen darin, dass die Datentransformation analysespezifisch im Rahmen der eigentlichen Modellbildung nach dem Laden erfolgt. Beide Verfahren haben große Überschneidungen angewandter Mechanismen und Technologien, sodass in diesem Dokument eine Beschränkung auf ETL-Prozesse erfolgt (siehe Bild 1).

Um die Datenqualität in Bezug auf die gestellten Zielvorgaben zu überprüfen, sind zusätzlich Funktionen zur Visualisierung und zur statistischen Analyse notwendig. Mit dieser ersten Datenanalyse erkannte Datenprobleme können dann eventuell wiederum in den ETL-Prozessen bereinigt werden.

Zu beachten ist, dass die gleichen ETL-Prozesse, die zur Modellbildung herangezogen wurden, auch bei der Modellanwendung durchlaufen werden müssen. Aus diesem Grund werden sowohl die genutzten Prozesse als auch die ermittelten Modelle während der Modellbildung in einem Repository abgelegt und bei der Modellanwendung genutzt.

Sofern mit vertretbarem Aufwand möglich, sollten schon an der Datenquelle Maßnahmen ergriffen werden, um die Datenqualität zu gewährleisten. Dies kann z.B. durch Masse- und Energiebilanzen oder durch redundante Messeinrichtungen erfolgen. Häufig sind solche Maßnahmen aber nicht mit vertretbaren Kosten realisierbar, sodass die Datenbereinigung im ETL-Prozess oder im Rahmen der Datenanalyse erfolgen muss.

Für den Einblick in die Thematik erklärt die Richtlinie zuerst die Datenstrukturformate, die heute an

The technical processes required for this are called ETL processes, where ETL stands for *extract*, *transform*, and *load*. Data management architectures of modern big data applications often provide for ELT (*extract, load, transform*) processes. These differ from ETL processes in that data transformation is analysis-specific and takes place as part of the actual model building process after loading. Both processes have large overlaps of applied mechanisms and technologies, so that in this document a restriction is made to ETL processes (see Figure 1).

In order to check the data quality with respect to the set targets, additional functions for visualization and statistical analysis are necessary. Data problems detected with this initial data analysis can then possibly be cleaned up in turn in the ETL processes.

It should be noted that the same ETL processes that were used to build the model shall also be run through in the model application. For this reason, both the processes used and the models determined are stored in a repository during model building and used during model application.

If possible with reasonable effort, measures should already be taken at the data source to ensure data quality. This can be done, for example, by mass and energy balances or by redundant measuring equipment. Often, however, such measures cannot be implemented at reasonable cost, so that data cleansing shall take place in the ETL process or as part of the data analysis.

For insight, the standard first explains the data structure formats that are encountered today (Section 4),

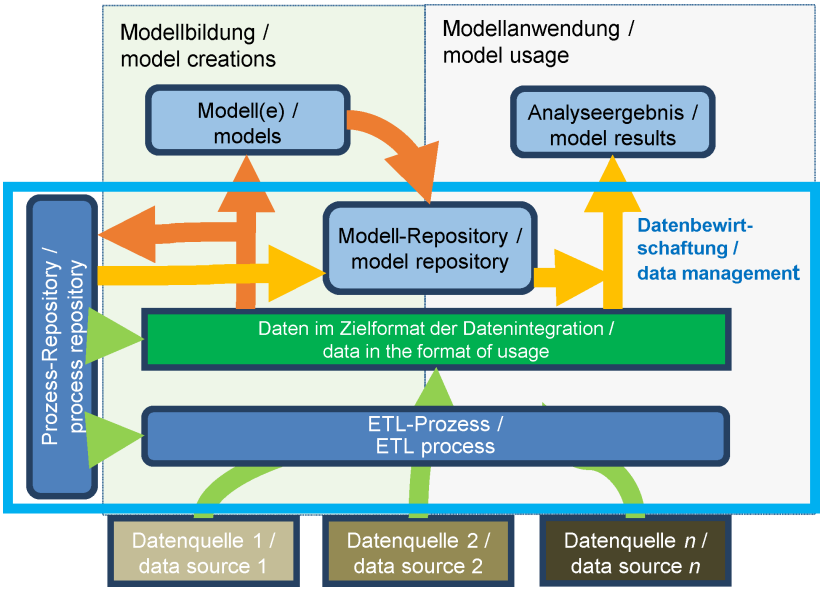

Bild 1. Übersicht über die Datenbewirtschaftung / Figure 1. Overview of data management

zutreffen sind (Abschnitt 4), erläutert dann sehr ausführlich Aspekte der ETL-Prozesse (Abschnitt 5), die in einem Big-Data-Projekt zu beachten sind und häufig zu Erfolgen in solchen Projekten beitragen. Außerdem wird darauf hingewiesen, wie wichtig es ist, Quellsysteme durch Datenextraktion nicht so zu belasten, dass sie ihre operativen Aufgaben nicht mehr mit der nötigen Performanz erfüllen können (Abschnitt 6). Abschnitt 7 weist auf derzeit verfügbare Tools im Markt hin, ohne Anspruch auf Vollständigkeit zu erheben.

Abschnitt 8 erklärt für die klassischen Produktionsverfahren „Stückgutfertigung“, „Batchherstellung“ und „kontinuierliche Prozesse“, wie bei den jeweiligen Verfahren Zusammenhänge zwischen den Daten hergestellt werden können.

Die Richtlinie schließt mit einem kurzen Überblick (Abschnitt 9), wie erste Analysen durchzuführen sind, um mangelhafte Daten zu erkennen und entsprechende Maßnahmen zu implementieren.

1 Anwendungsbereich

Mit dem Begriff „Big Data“ werden – obwohl er bereits seit einigen Jahren verwendet wird – unverändert sehr unterschiedliche Themen und Aspekte assoziiert und entsprechend in der gesellschaftlichen Diskussion differenziert diskutiert. Die immer weiter voranschreitende digitale Kommunikation, der in der Umsetzung befindliche Breitbandausbau und die überall mögliche Verarbeitungsmöglichkeit von Daten beflügeln diese Diskussion sowohl in der Öffentlichkeit als auch in der Fachwelt. Die Themen reichen von Datenschutz und Datensicherheit bis hin zu generellen Strategien für die digitale Wertschöpfung bei kleinen und mittelständischen Unternehmen und auch bei Großunternehmen.

Im Kontext dieser Richtlinie geht es bei Big Data um Technologien zur Datenanalyse. Entsprechende Algorithmen und Werkzeuge können Erkenntnisse über betriebliche Abläufe liefern und zu deren Optimierung beitragen. Hierzu bedarf es der Umsetzung dieser Methoden und Werkzeuge zur Verarbeitung, Analyse und Interpretation von umfangreichen und komplexen Daten in Big-Data-Anwendungen. Die Richtlinienreihe unterstützt Erstellende und Nutzende bei der Vorbereitung, Entwicklung und Inbetriebnahme dieser Anwendungen sowie deren nachhaltigen Einsatz. Letztlich sollen diese Big-Data-Anwendungen verlässlichere Entscheidungsgrundlagen schaffen, um Produkte und Produktionsprozesse ökonomisch, ökologisch und technisch zu verbessern.

Die Richtlinienreihe soll dazu beitragen, die Vielfalt der in den letzten Jahren durch Forschungs-,

then explains in great detail aspects of ETL processes (Section 5) that need to be considered in a big data project and often contribute to successes in such projects. It also points out the importance of not overloading source systems with data extraction in such a way that they can no longer perform their operational tasks with the necessary performance (Section 6). Section 7 points out tools currently available on the market without claiming to be exhaustive.

Section 8 explains for the classic production processes “piece goods production”, “batch production” and “continuous processes” how correlations between data can be established for the respective processes.

The standard ends with a brief overview (Section 9) of how initial analyses are to be carried out in order to identify defective data and to implement appropriate measures.

1 Scope

Although the term “big data” has been in use for several years, it continues to be associated with a wide variety of topics and aspects and is accordingly the subject of differentiated discussion in society. The ever-advancing digital communication, the broadband expansion that is currently being implemented, and the processing of data that is possible everywhere are fuelling this discussion both among the public and among experts. The topics range from data protection and data security to general strategies for digital value creation for small and medium-sized enterprises as well as for large companies.

In the context of this standard, big data is about data analysis technologies. Corresponding algorithms and tools can provide insights into operational processes and contribute to their optimization. This requires the implementation of these methods and tools for processing, analysing and interpreting extensive and complex data in big data applications. The series of standards supports creators and users in the preparation, development, and commissioning of these applications as well as their sustainable use. Ultimately, these big data applications should create a more reliable basis for decision-making in order to improve products and production processes economically, ecologically, and technically.

The series of standards is intended to help process the wide range of findings that have emerged in

Entwicklungs- und Praxisarbeiten entstandenen Erkenntnisse aufzubereiten, die Entwicklung und den Einsatz von Big-Data-Anwendungen in produzierenden Industrien sowie deren Nutzung im regulären Betrieb zu unterstützen.

Zur Zielgruppe gehören alle Stakeholder, von den Praktikern/Praktikerinnen bis zu den Entscheidern/Entscheiderinnen, von der Fertigungs- bis zur Prozessindustrie. Die Richtlinienreihe wendet sich dabei an die Nutzenden und die Erstellenden von Big-Data-Anwendungen in der produzierenden Industrie, unabhängig und übergreifend für alle Führungs- und Fachaufgaben.

Diese Richtlinie behandelt den Vorgang, wie vorhandene Daten aus der Produktion einer Analyse zugänglich gemacht werden können. In den meisten Fällen ist es notwendig, die Daten aus verschiedenen Datenquellen zusammenzuführen und in einer geeigneten Weise so zu transformieren, dass eine Datenanalyse überhaupt erst möglich wird.

Die Richtlinie soll Big-Data-Interessierten von der Fertigungs- bis zur Prozessindustrie einen leicht verständlichen Einblick in die Thematik und einen Überblick über die Datenbewirtschaftung geben. Dazu gehören die notwendigen ETL-Verfahren und die erste Analyse der Daten, denn häufig sind Iterationen notwendig, ehe die Daten, die von den Quellen extrahiert und aufbereitet worden sind, den Anforderungen der weiteren Verwendung genügen.

recent years through research, development, and practical work, and to support the development and use of big data applications in manufacturing industries as well as their use in regular operations.

The target audience includes all stakeholders, from practitioners to decision makers, from operations to process industries. In this regard, the series of standards addresses the users and the creators of big data applications in the manufacturing industry, independently and across all management and technical tasks.

This standard covers the process of making existing data from production accessible for analysis. In most cases, it is necessary to bring together data from different data sources and transform it in a suitable way so that data analysis becomes possible in the first place.

The standard is intended to give those interested in big data from the operation to the process industry an easy-to-understand insight into the subject and an overview of data management. This includes the necessary ETL procedures and initial analysis of the data, as iterations are often necessary before the data that has been extracted and processed from the sources meets the requirements for further use.