

VEREIN
DEUTSCHER
INGENIEURE

VERBAND DER
ELEKTROTECHNIK
ELEKTRONIK
INFORMATIONSTECHNIK

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

VDI/VDE 3714

Blatt 3

Entwurf

Implementation and operation of big data applications in the manufacturing industry – Data management

Einsprüche bis 2020-05-31

- vorzugsweise über das VDI-Richtlinien-Einspruchsportal <http://www.vdi.de/3714-3>
- in Papierform an
VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik
Fachbereich Digitale Transformation
Postfach 10 11 39
40002 Düsseldorf

Inhalt	Seite
Vorbemerkung	2
Einleitung	2
1 Anwendungsbereich	4
2 Normative Verweise	4
3 Begriffe	4
4 Generelle Datenstrukturformate	4
4.1 Strukturierte Daten	5
4.2 Semistrukturierte Daten	5
4.3 Unstrukturierte Daten	5
4.4 Polystrukturierte Daten	5
5 ETL-Prozesse	5
5.1 Extraktionsprozesse	6
5.2 Transformationsprozesse und Datenbereinigung	8
5.3 Ladeprozesse	10
6 Quellenverfügbarkeit und -belastung	11
7 Einsatz von ETL-Tools	12
8 Typische Datenzusammenhänge bei spezifischen Prozessarten	12
8.1 Fertigung von Stückgut	13
8.2 Batchprozesse	13
8.3 Kontinuierliche Anlagen	14
9 Erste Analyse der Daten	15
9.1 Statistische Beurteilung der Daten	16
9.2 Mögliche Datenbereinigung	17
Schrifttum	20

VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA)

Fachbereich Digitale Transformation

VDI/VDE-Handbuch Automatisierungstechnik

Vorbemerkung

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

Alle Rechte, insbesondere die des Nachdrucks, der Fotokopie, der elektronischen Verwendung und der Übersetzung, jeweils auszugsweise oder vollständig, sind vorbehalten.

Die Nutzung dieser Richtlinie ist unter Wahrung des Urheberrechts und unter Beachtung der Lizenzbedingungen (www.vdi.de/richtlinien), die in den VDI-Merkblättern geregelt sind, möglich.

An der Erarbeitung dieser Richtlinie waren beteiligt:

Assoc. Prof. Dr. habil. *Martin Atzmüller*, Tilburg

Dipl.-Ing. *Sven Behrendt*, Worms

Dr. Dipl.-Chem. *Raphael Benner*, Heilbronn

Dipl.-Inf. *Rolf Blumenthal*, Lüneburg

Dr.-Ing. *Markus Feist*, Ludwigshafen

Dr. rer. nat. *Eric Frauendorfer*, Burghausen

Dipl.-Ing. *Thomas Froese*, Willich

Prof. Dr. *Jochen Garcke*, Sankt Augustin

Dr. *Jörg Gebhardt*, Hahnstätten

Dr.-Ing. *Ingo Gräf*, Frankfurt

Dipl.-Ing. *Herbert Grieb*, Karlsruhe

Dr.-Ing. *Gunnar Hoffmann*, Essen

Jakob Kinghorst M.Sc., Böblingen

Dipl.-Ing. *Christoph Kugler*, Würzburg

Dr. Dipl.-Chem. *Heiko Kulinna*, Lampertheim

Dr.-Ing. *Silke Müller*, Ratingen

Dr.-Ing. *Dirk Ortloff*, Peterhausen

Dipl.-Math. *Roland Rosen*, München

Prof. Dr.-Ing. *Heiko Tapken*, Osnabrück

Hasan Tercan M.Sc., Aachen

Iris Weiß M.Sc., München

Prof. Dr.-Ing. *Winfried Wilke*, Schweinfurt

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

Einleitung

Der Fachausschuss „Big Data“ der VDI-Gesellschaft für Mess- und Automatisierungstechnik hat sich mit der Erstellung dieser Richtlinie der Aufgabe angenommen, den ökonomischen und ökologischen Nutzens 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 bzw. welche Einschränkungen und Hindernisse bestehen. Dem Praktiker 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
(in Vorbereitung)

Blatt 5 Modellierungsverfahren (in Vorbereitung)

Blatt 6 Validierung von Modellen
(in Vorbereitung)

Blatt 7 Online-Anwendung von datengetriebenen Modellen (in Vorbereitung)

Eine Liste der aktuell verfügbaren Blätter dieser Richtlinienreihe ist im Internet abrufbar unter www.vdi.de/3714.

Die Richtlinienreihe VDI/VDE 3714 ist im Fachausschuss 7.24 „Big Data“ des Fachbereichs 7 „Anwendungsfelder der Automation“ der VDI/VDE-Gesellschaft für 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.

Benachbart zu Big Data finden sich Themen wie das Internet der Dinge (IoT), 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.

Die Richtlinienreihe geht von einer generellen Verfügbarkeit aller benötigten Daten aus. Bezüglich Datenmenge, ihrer Struktur und Integrität wird keine Annahme getroffen. Zur Diskussion und

Charakterisierung der Daten helfen die fünf sogenannten „V“, die die einzelnen Dimensionen von Big Data bezeichnen. Die Daten werden durch Umfang (*Volume*), Unterschiedlichkeit (*Variety*) und ihre Schnellebigkeit (*Velocity*) charakterisiert. Insbesondere bei industriellen Anwendungen sind die Qualität der Daten (*Validity*) und der unternehmerische Mehrwert (*Value*) relevant.

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 900x
- 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.

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 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.

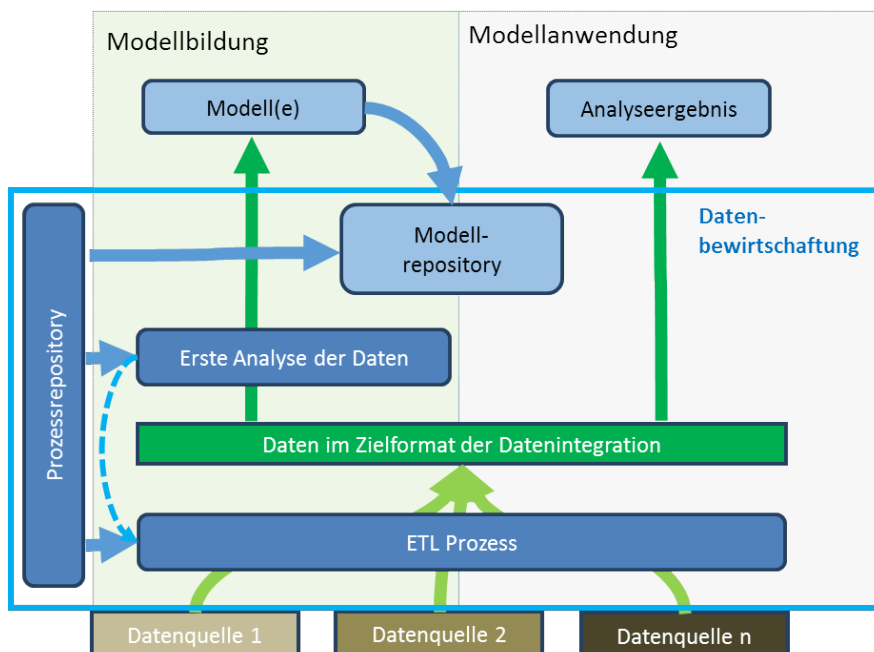


Bild 1. Übersicht über die Datenbewirtschaftung

Für den Einblick in die Thematik erklärt die Richtlinie zuerst die Datenstrukturformate, die heute anzutreffen 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 einzubauen.

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 klein- und mittelständischen Unternehmen und auch für 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 Ersteller und Nutzer bei der Vorbereitung, Entwicklung und Inbetriebnahme dieser Anwendungen sowie ihrem nachhaltigen Einsatz. Letztendlich 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-, 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 bis zu den Entscheidern, von der Fertigungs- bis zur Prozessindustrie. Die Richtlinienreihe wendet sich dabei an die Nutzer und die Ersteller 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.

2 Normative Verweise

Das folgende zitierte Dokument ist für die Anwendung dieser Richtlinie erforderlich:

VDI 3714 Blatt 1:2019-12 Implementierung und Betrieb von Big-Data-Anwendungen in der produzierenden Industrie; Durchführung von Big-Data-Projekten