Eine nationale Daten- und Analyseinfrastruktur als Grundlage Digitaler Souveränität

Zusammenfassung

Daten sind der wesentliche Produktionsfaktor der Zukunft. Informatik und Data Science werden zur Grundlage der meisten anderen wissenschaftlichen Fächer wie Physik, Chemie, Wirtschaft, sie sind in ihrer Bedeutung damit der Mathematik vergleichbar. Datenzugang und Datenkompetenz sind die Schlüsselfaktoren für die zukünftige Wettbewerbsfähigkeit. In beiden Belangen ist Deutschland ins Hintertreffen geraten. Zur Erlangung von Datensouveränität benötigen Deutschland bzw. Europa eine nachhaltig (und öffentlich) betriebene Daten- und Analyseinfrastruktur. Diese stellt den Zugang zu Daten zu großen, qualitativ hochwertigen Datenmengen (Internet, Forschungsdaten, öffentliche Daten z.B. mCloud) sicher und ermöglicht es, deren Analyse und Visualisierung für Schulen, Universitäten, Forschungseinrichtungen und Bürger zu demokratisieren. Nur auf diese Weise lässt sich die Lücke zu den monopolisierten Datensammlungen amerikanischer IT Konzerne oder dem staatlich regulierten Zugang zu Daten in China schließen. Eine derartige nationale, allgemein zugängliche Infrastruktur sollte nicht nur Daten, Analysen und Visualisierungen verwalten und kontinuierlich in Echtzeit bereitstellen, sondern gleichzeitig die Algorithmen und Werkzeuge entlang der gesamten Datenwertschöpfungskette (Quellenauswahl, Informationsextraktion und Integration, Analyse und Modellbildung sowie Anwendung und Visualisierung) einfach nutzbar (web-basiert, open-source, wiederverwendbar) bereitstellen, um auf diese Weise durch „Daten und Analysen aus der Steckdose“ Forschung sowie die Innovation in datengetriebene Anwendungen in Wirtschaft, Wissenschaft und Gesellschaft befeuern.

Hintergrund, Herausforderungen und Potentiale

Die Vielfalt und die exponentiell wachsende Menge digitaler Daten, die heute in Wirtschaft, Medizin, Mobilität sowie in vielen weiteren Lebensbereichen anfallen, bieten zusammen mit neuen Verfahren des maschinellen Lernens und der künstlichen Intelligenz gänzlich neue Chancen, automatisiert neue Muster und Zusammenhänge zu erkennen, Frühwarnungen zu erzeugen oder Prozesse zu steuern. Maschinen können – im Gegensatz zu Menschen – aus großen Datenströmen in Echtzeit lernen, sie helfen uns, die neue Flut an Daten und Informationen zu beherrschen und produktiv zu nutzen. Dabei spielen Menge und Qualität der Daten über die Möglichkeiten und Mächtigkeit der maschinellen Lern- und KI-Verfahren und Anwendungen eine entscheidende Rolle.

Dauer und Kosten von Analyseprozessen sind dank der neuen Datenquellen und neuer KI-Verfahren für viele Berufe, für die Wissenschaft und den Bürger drastisch gesunken. Relevante Inhalte können schneller gefunden und wichtige Zusammenhänge schneller erkannt werden. Die Geschwindigkeit der Wissenserzeugung aus Daten ist heute mehr denn je zum entscheidenden Wettbewerbsvorteil geworden.

Wenn Daten die Produktionsfaktoren der Zukunft sind, dann sind Datenzugang und Datenanalysekompetenz die Schlüsselfaktoren für zukünftige Wettbewerbsfähigkeit. In beiden Belangen ist Deutschland leider im internationalen Vergleich ins Hintertreffen geraten.

Deutschland nimmt zwar nach wie vor eine Spitzenposition in der (wissenschaftlichen) Technologieforschung ein, jedoch sind es derzeit vor allem US-amerikanische und zunehmend auch

asiatische Unternehmen, die wichtige Daten- und Analyseinfrastrukturen für die neue digitale Wirtschaft bereitstellen. Die Angebote der großen Software as a Service (SaaS) Anbieter Amazon Web Services, Google Data Flow, Microsoft Azure, IBM Bluemix dominieren aufgrund ihrer überlegenen Funktionalität, Performance und Usability sowohl in wissenschaftlichen wie auch wirtschaftlichen Anwendungen. Entsprechend machen sich europäische Akteure immer stärker von diesen kostenpflichtigen Angeboten abhängig. Und nicht nur das: in der Folge liegen die meisten europäischen Daten, Konsumentendaten wie auch Unternehmensdaten, außerhalb Europas und werden von Software nichteuropäischer Unternehmen in Drittländern analysiert. Wichtigste Voraussetzung für digitale Souveränität ist jedoch uneingeschränkte Verfügungsgewalt über unsere Daten und den Quellcode der Analysesoftware sowie die Durchführung der Analysen auch geographisch in unserem Rechtsrahmen.

Dem Markt gelingt es nicht, eine unabhängige deutsche oder europäische Alternative mit einem eigenen, gleichwertigen Angebot zu entwickeln, obwohl das technologische Potenzial dafür durchaus vorhanden ist, wie Beispiele international anerkannter und eingesetzter Technologieplattformen wie Apache Flink [1], Rapidminer [2} und OpenML [3] zeigen. Grund ist, dass eine vergleichbare kritische Masse hier von keinem Anbieter erreicht werden kann und in Deutschland insbesondere auch die Kompetenz zum Aufbau großer Daten- und Analyseinfrastrukturen bei keinem Unternehmen vollumfänglich vorhanden ist. Um den Leistungsabstand zu den dominierenden Plattformen zu verkürzen und eine eigenständige deutsche oder europäische Entwicklung mit dem damit verbundenen Kompetenzaufbau bei Erstellung und kontinuierlichem Betrieb von großen Daten- und Analyseinfrastrukturen anzustoßen, wird ein Anschub benötigt. Den Anfang für ein solches Angebot könnte eine nachhaltig (und öffentlich) betriebene Datenanalyseinfrastruktur sein, die im Gegensatz zu den amerikanischen Angeboten weniger auf der Auswertung und Vermarktung von Konsumdaten basiert, sondern vielmehr breitere, für die europäische Wirtschaft und Gesellschaft relevantere Domänen, wie bspw. Mobilität und Gesundheit, in den Fokus stellt und gleichzeitig den Zugang zu großen, qualitativ hochwertigen Datenmengen (Internet, Forschungsdaten, öffentliche Daten) und deren Analyse und Visualisierung in Echtzeit für Unternehmen, Universitäten, Forschungseinrichtungen, Schulen und Bürger ermöglicht und demokratisiert.

Damit entsteht eine ausreichend große „Spielwiese“, ein Nährboden für datengetriebene Technologie-Innovationen und Unternehmensgründung entlang der Wertschöpfungskette der Daten- und Analyseinfrastruktur, von Datenquellenauswahl über Veredelung der Rohdaten durch Informationsextraktion und Integration bis hin zur Erstellung von Datenanalysen und Modellbildung durch Verfahren des maschinellen Lernens und der künstlichen Intelligenz. Voraussetzung ist die auf Dauer angelegte Etablierung und der Betrieb dieser Infrastruktur. Heutige Forschungsprogramme sind zumeist auf drei Jahre angelegt, in der einzelne Projekte mühsam eine eigene Daten- und Analyseinfrastruktur aufbauen, Daten kurieren und analysieren, um nach Projektabschluss den Server und die Dienste abzuschalten und die gewonnenen wertvollen Daten zu verwerfen. Dieser Prozess erzeugt nicht nur erheblichen Mehrfachaufwand, er verschenkt auch das Potenzial, durch die kontinuierliche Arbeit an einer gemeinsamen Daten- und Analyseinfrastruktur große Datenbestände aufzubauen, Analyse-Werkzeuge kontinuierlich weiterzuentwickeln und alle Projektergebnisse inklusive Daten und Code nachhaltig auch nach dem Ende eines Projektes für die Allgemeinheit nutzbar zu machen.

Eine *logisch* zentrale, nationale, allgemein zugängliche Infrastruktur könnte dagegen nicht nur eine große Vielzahl und Vielfalt an Daten kontinuierlich bereitstellen, sondern gleichzeitig Werkzeuge der gesamten Datenwertschöpfungskette (Aufbereitung, Analyse und Visualisierung) einfach nutzbar (web-basiert, plug&play, Kombination von öffentlichen und privaten Daten in einer Analyse) anbieten und über ihre Nutzung stetig weiterentwickeln. Eine derartige Infrastruktur kann nicht nur von

Akteuren in der Wirtschaft genutzt werden, sondern auch von Forschungseinrichtungen, Universitäten, Schulen und Bürgern insgesamt. Auf diese Weise kann ein derartiges System ein Innovationmotor für Ausbildung sowie datenorientierte Wertschöpfung, Geschäftsmodell-innovationen und Unternehmensgründungen werden.

Ein wesentlicher Vorteil einer öffentlich betriebenen Infrastruktur wäre die konsequente Umsetzung offener Standards, mittels derer der derzeit gängige Vendor-Lock-in für Daten- und Analyseinfrastrukturen gebrochen werden könnte. Damit könnte eine nationale bzw. europäische Datensouveränität erreicht und die Abhängigkeit von Anbietern aus Drittländern beim in der Zukunft immer wichtigeren Produktionsfaktor Daten verringern oder ganz abgebaut werden.

Im Bereich der Wissenschaft ist als ein bereits existierender Anwendungsfall einer derartigen Infrastruktur der Sloan Digital Sky Server (SDSS) [4] zu nennen, der als domänenspezifische Infrastruktur mit Daten und Analysefunktionen in einem System bereits seit Jahren in Betrieb ist und vor seiner Einführung auch von einigen Skeptikern als unmöglich angesehen wurde, aufgrund von Datenvolumen, Anfragekomplexität, etc. Wesentliche Protagonisten waren der leider inzwischen verstorbene Turing-Preisträger Jim Gray, der den Begriff des 4. Paradigmas der Wissenschaft [5] geprägt hat. Der SDSS hat zu vielen spannenden wissenschaftlichen Erfolgen geführt, neben der Nutzung durch die wiss. Community auch zum spielerischen Umgang mit den Daten durch Outsider und Hobbyisten, was u.a. in dem Auffinden von neuen Galaxien und anderer Phänomene durch interessierte Laien („Citizen Science“) gipfelte. Aufbauend auf den Erfahrungen mit SDSS entsteht derzeit in den USA unter Federführung der Johns Hopkins Universität, Prof. Alex Szalay, ein Big Data Open Storage Network [6], das die logisch zentrale Datenverwaltung für einen Großteil der öffentlich geförderten Forschungsprojekte übernehmen soll.

Bezüglich wirtschaftlicher Anwendungen ist zu befürchten, dass die meisten großen deutschen Unternehmen den Wert einer nationalen Daten- und Analyseinfrastruktur nicht sofort erkennen werden. Ich würde hier (leider) im Vergleich zum angelsächsischen und nordeuropäischen Raum eine langsamere Adoption erwarten, analog zu der verspäteten Adoption von Internet, Big Data, KI-Methoden, Cloud, etc.

Es besteht jedoch riesiges Potential für eine erste Nutzung bei „neuen“ Unternehmen, also Startups und jungen Unternehmen, bei denen Daten den wesentlichen Teil des Kerngeschäfts darstellen bzw. massive Wettbewerbsvorteile liefern, z.B. bei Marktforschungsdaten bzw. im Bereich Mobilität, Logistik, Handel. Diese könnten durch Skaleneffekte schneller am Markt aktiv werden, ohne eine Daten- und Analyseinfrastruktur selbst aufzubauen und unterhalten zu müssen. Die Erwartung ist, dass sich aufgrund von Startups durch schnelle Umsetzung in diesen Bereichen durch Positivbeispiele die Innovationszyklen und Technologieadoption auch in den langsameren großen Industrieunternehmen verkürzen. Ferner, dass die Technologien auch in den Mittelstand wirken, aufgrund einer geringeren Eintrittsschwelle bei der Nutzung einer Daten- und Analyseinfrastruktur. Außerdem würde eine derartige Infrastruktur die Wiederverwendung von Algorithmen und Daten in Wirtschaft und Wissenschaft fördern und somit die Kosten für neue Erkenntnisse bzw. die Etablierung von neuen Geschäftsmodellen senken. Gleichzeitig ist ein großer gesamtgesellschaftlicher Nutzen im Hinblick auf Data Literacy für die Bevölkerung insgesamt zu erwarten, da Schüler auf diese Weise, z.B. bei der Vorbereitung von Referaten oder Hausarbeiten, spielerisch an Programmierung, Datenanalyse und sogar potentielle Geschäftsmodelle oder „Citizen Science“ herangeführt werden könnten, indem sie „Apps“ on top of der Daten- und Analyseinfrastruktur entwickeln.

Anforderungen an eine solche Daten- und Analyseinfrastruktur

Vom Vorbild kommerzieller Infrastrukturen mit Fokus auf Datenverarbeitung können folgende konkrete Anforderungen für eine derartige Infrastruktur abgeleitet werden:

1. Geringe Einstiegshürde für die Erstellung und Nutzung von Datenanalysen, Visualisierungen, Modellen durch deklarative Sprachen, in denen lediglich gewünschte Ergebnis beschrieben werden und nicht die Algorithmen zu dessen berechnen, gepaart mit visuellen Tools
2. Ubiquitärer Zugang zur Datenanalyseinfrastruktur durch einen web-browser-basierten Zugang durch Mobiltelefonen, Laptops, etc. ohne die Anforderung, Software installieren zu müssen
3. Nutzung und Wiederverwendung des Wissens der Community durch Verfügbarmachen der Extraktions-, Integrations-, Analyse- und Visualisierungsmethoden in einem Repository, mit Bewertungen und Nutzungsbeispielen von Daten und Verfahren
4. Zugriff auf Datenströme in Echtzeit und die Möglichkeit, kontinuierliche Analyse auf einem oder mehreren Datenströmen gepaart mit anderen Datenquellen zu realisieren
5. Verantwortungsvolles Datenmanagement durch Sicherstellung von Daten(analsye)schutz durch Zweckbindung, Nachvollziehbarkeit und Reproduzierbarkeit der Analyseergebnisse

Daraus leiten sich diverse Anforderungen an das technische System ab:

1. Benutzerfreundlichkeit, Funktionalität und Kosteneffizienz (deklarative Programmier-schnittstellen und automatische Optimierung, modularer Aufbau)
2. Breite Verfügbarkeit von Quellen mit guter Qualität der Daten und Datenströme (Open Data, Datenspenden, Möglichkeit der Speicherung von Daten mit Zugriffsrestriktionen, Bereinigung, Kuratierung, Integration)
3. Skalierbarkeit des Systems bezüglich Datenmenge, Datenvielfalt und hoher Rate des Zugangs neuer Daten, Nutzerzahl und Analysekomplexität (durch Aufbau einer Rechnerarchitektur basierend auf moderner Hardware)
4. Community-Funktionen zum Teilen und Bewerten von Daten, Analysemethoden, Visualisierungen, durch Open-Source Entwicklung, dokumentierte Open-Source APIs, Offenheit
5. Vertrauenswürdigkeit, Daten(analyse)schutz (Sicherstellung der Zweckbindung der Analysen), Zugriffsrechte, Sicherheit, Transparenz, Nachvollziehbarkeit, Reproduzierbarkeit
6. Unterstützung des gesamten Datenlebenszyklus sowie der vollständigen, oft iterativen Verarbeitungsketten des maschinellen Lernens, von Quellauswahl über Datenaufbereitung über Analyse zur Modellanwendung und Visualisierung durch verschiedene, einander überführbare Programmierabstraktionen, von visuellen Tools über deklarativen Sprachen zu imperativen Code der Algorithmen, welcher parallelisiert, verteilt und auf der Datenanalyseinfrastruktur optimiert zur Ausführung gebracht wird
7. Aufbau einer logisch zentralen, massiv-parallelen, physisch verteilten, multi-tenancy Hardwareinfrastruktur zur Speicherung und kontinuierlichen Erweiterung der Daten sowie Verarbeitung der Datenströme und Analysealgorithmen durch intelligentes Caching mit Speicherung und Wiedernutzung bereits durchgeführter Analysen im Rahmen von neuen Analysen

Der Erfolg einer Daten- und Analyseinfrastruktur hängt vor allem von der Funktionalität, der Benutzerfreundlichkeit bei der Erstellung von Analysen und dem Datenangebot ab. Solange ausländische, kommerzielle Anbieter vergleichbare Funktionalität anbieten können und gleichzeitig dieselben Daten zur Verfügung stellen, wird eine neue Plattform nicht angenommen werde.

Um eine Nutzbarkeit von verfügbaren Daten zu ermöglichen ist es wichtig Daten und Ströme nicht nur in ihrer Rohform zur Verfügung zu stellen, sondern auch Metadaten zu verwalten und sogar

automatisch abzuleiten und kontinuierlich zu verfeinern, zum Beispiel durch Informationsextraktion (z.B. Erkennung von Ereignissen aus Textdatenquellen) oder fortgeschrittene Analyse (z.B. Erkennung von zeitabhängigen Mustern bei der Auslastung von Verkehrsmitteln). Durch die Integration und Abbildung von verschiedenen Datenquellen auf dieser Plattform können individuelle Datensätze nach Bedarf angereichert und erweitert werden. Dazu gehören auch Verfahren zur rechtsicheren Anonymisierung und Pseudonymisierung personenbezogener Daten, wie sie bspw. bei medizinischen Analysen gefordert sind.

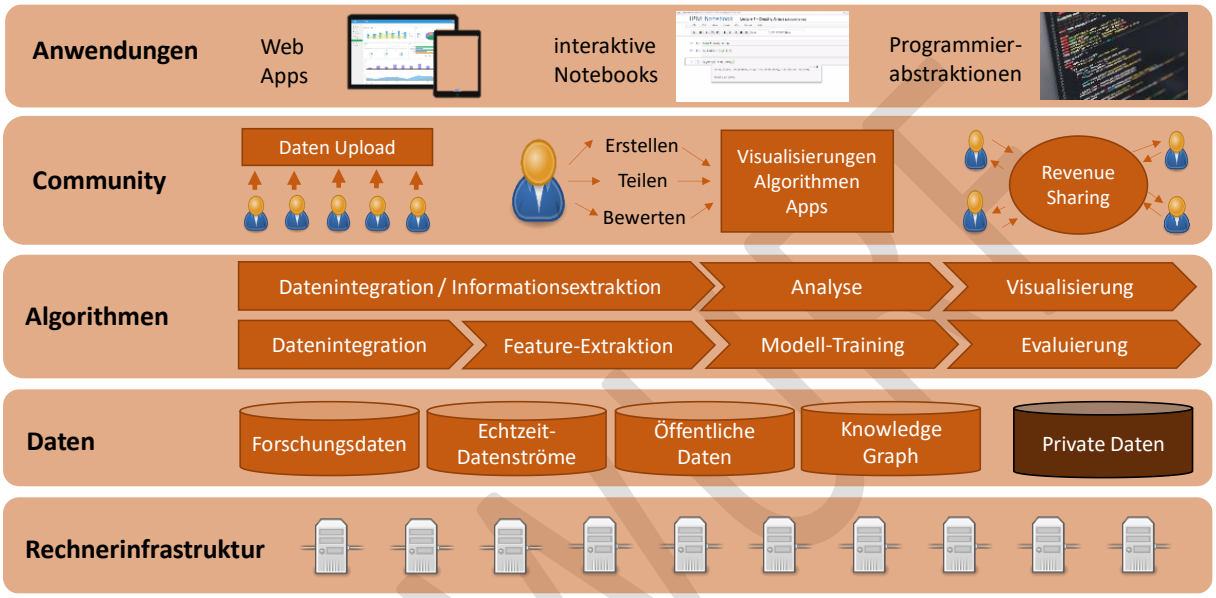
Auf dieser Infrastruktur könnten dann kontinuierlich wachsende Datenmengen sowie Qualitätssicherungs- und Analyseverfahren entwickelt werden, die domänenspezifische Anwendungen unterstützen und durch eine wachsende Anzahl an einfach zu nutzenden, von einer open-source community entwickelten und auf der Infrastruktur bereitgestellten Data Science Werkzeugen weiterverarbeitet werden. Eine wichtige Datenquelle stellen kontinuierliche Datenströme in Echtzeit dar, die es möglich machen, neue Erkenntnisse zum Zeitpunkt der Datenerhebung zu gewinnen. Die Entwicklung deklarativer Konzepte, die es auch Menschen ohne ausgeprägte Informatikkenntnisse ermöglichen sollen, Extraktions-, Integrations-, Analyse-, Stromverarbeitungs- und Informationsvisualisierungsmethoden in einem Baukastenprinzip anzuwenden, ist der Schlüssel zur breiten Akzeptanz und damit letztlich erfolgreichen Implementierung nachhaltiger digitaler Informationsinfrastrukturen.

Die rasante Entwicklung der Hardware-Landschaft in den letzten 10 Jahren erweitert hierbei die Datenverarbeitungsmöglichkeiten, wichtige Technologien sind hierbei Multi-Core-CPUs, GPUs sowie spezielle Beschleuniger wie Xeon Phi Coprozessoren, FPGAs und TPUs. Die heutige Hardware-Landschaft erfordert, dass Daten- und Analyseinfrastrukturen verschiedene Prozessoren und folglich verschiedene Programmiermodelle unterstützen, um die Leistungsanforderungen von Benutzern zu erfüllen. Eine große Herausforderung in diesem Bereich ist eine gemeinsame Abstraktion für Daten- und Analyseinfrastrukturen, mit der Nutzer die Algorithmen mit heterogenen Prozessoren und Beschleunigern ohne großen Aufwand zur Ausführung bringen können. Weiterhin ist eine Unterstützung neuer Speichertechnologien wie SSDs und NVRAM (nicht flüchtiger Hauptspeicher) erforderlich, um auf große Datenmengen schnell zugreifen zu können und echtzeitnahe Analyse zu ermöglichen.

Eine erfolgreiche Daten- und Analyseinfrastruktur wird dem Benutzer alle Stufen der Datenverarbeitungskette des maschinellen Lernens anbieten und optimieren, dazu zählen Datenerhebung, Datenvorverarbeitung, Feature Extraction, Modellbildung und Training, Modellinstallation und -update, sowie die Inferenz. Die einfache Durchführung von Analysen im Web-Browser und die Verhinderung von Abhängigkeiten durch community-basierte open-source Entwicklung und Bereitstellung von Daten und Verarbeitungsketten sollte dabei ein entscheidender Vorteil einer öffentlichen, deutschen oder europäischen Lösung sein.

Die hier skizzierte Daten- und Analyseinfrastruktur kann auf eine Vielzahl an grundlegenden Vorarbeiten aufbauen, sowohl im Bereich der Forschungsdateninfrastrukturen [7,8,9], der Wirtschaft (z.B. dem International Data Space[10]) und der Vorarbeiten im Kompetenzzentrum ScaDS und BBDC [11,12]. Eine derartige Infrastruktur unterscheidet sich jedoch von klassischen Höchstleistungsrechnern, in Bezug auf Fokus von effizienter Datenübertragung und Datendurchsatz im Gegensatz zu Höchstleistungs-Prozessoren sowie in Bezug auf Software- und Hardware-Co-design mit Fokus auf Datenmanagement und Analyse sowie Management von Analysealgorithmen und der damit verbundenen Entwickler- und Nutzer-Community. Ferner unterscheidet sich eine derartige Infrastruktur auch von Open Data Lösungen, die häufig im Bereich der Forschungsdateninfrastrukturen bzw. für öffentliche Daten geschaffen werden bzw. dem Industrial Data Space, da sie die gesamte Datenwertschöpfungskette und die damit verbundenen Werkzeuge und Programmiermodelle sowie

effiziente, ggf. parallele und verteile Ausführung der Datenanalyseprogramme beinhaltet. Ferner unterscheidet sich eine derartige Infrastruktur auch von klassischen Datenbanksystemen, da die Analysen durch fortgeschrittene Methoden der mathematischen Programmierung, der Signalverarbeitung und des maschinellen Lernens durch Verwendung von Iterationen und komplexen benutzerdefinierten Funktionen über die Konzepte der relationalen Algebra hinausgehen. Von existierenden Vorhaben wie dem Industrial Data Space oder dem Smart Data Innovation Lab unterscheidet sich eine derartige Infrastruktur durch ihre generelle Ausrichtung und damit verbundene Skalierbarkeit und durch die Zielgruppe (Wirtschaft, Wissenschaft und Bürger insgesamt) sowie Zugriffsmöglichkeit (web-basiert, interatkiv).



Dabei ist anzumerken, dass aufgrund der verteilten Erzeugung der Daten eine zentrale Datenspeicherung aller anfallenden Rohdaten aus Datenströmen, z.B. alle Sensordaten von selbstfahrenden Autos mit derzeitigen Technologien nicht sinnvoll ist. Hierbei gilt der Satz des ehemaligen CTOs von Teradata, Todd Walther: „Data is as elastic as brick wall“, d.h., die Daten können nicht jederzeit ad-hoc übertragen werden. In einer derartigen Architektur sollten Metadaten, Modelle sowie Algorithmen und weitere durch Analysen abgeleitete Daten logisch zentral verwaltet werden, wobei Daten ggf. physisch verteilt vorliegen und die Berechnung und Ableitung von Modellen sollte nahe an der Datenquelle durchgeführt werden. Bei großen, insbesondere kontinuierlichen Datenquellen kann dies durchaus föderiert, im Sinne des sogenannten „function shipping“ stattfinden, um „data shipping“ und Kommunikation nach Möglichkeit zu vermeiden bzw. zu reduzieren. Aus Latenzgründen sollten jedoch dabei je nach Netz- und Speicherkapazität so viele Daten „zentral“ bzw. nahe an der Analysedurchführung gespeichert werden, wie für Analysen nötig, ggf. mit Prefetching für avisierte Analysen von verteilten Daten und ggf. föderierter Vorverarbeitung. Gleichzeitig sollte analog zu üblichen Data Warehouse Infrastrukturen der Drill-Through auf die Rohdaten möglich sein, falls erforderlich bzw. die (Vor-)Analyse der Rohdaten nahe der Datenquelle. Wichtig für die breite Nutzung der Infrastruktur von Wissenschaftlern, Unternehmen, in Schulen und von Bürgern insgesamt ist jedoch, dass die Infrastruktur die Verarbeitung von interaktiven Analysen in vielen Fällen ermöglicht, d.h., dass Analyseergebnisse und Visualisierungen für viele Klassen von Analysen im Rahmen einer „Denkzeit“ / „nahezu-Echtzeit“ produziert werden und Programme und Daten geteilt und wiederverwendet werden können. Dies erfordert das zentrale Caching/Vorhalten von bei Analysen benötigten Daten und die Verbindung von Compute und Storage im Sinne einer gemeinsamen Architektur und Software-/Hardware-Co-Design, was im klassischen HPC-Bereich eher unüblich ist. Ferner muss die Infrastruktur alle Aspekte der Datenanalyse, von Informationsextraktion und -

integration/Datenkuration über Analyse und Modellbildung bis hin zur Visualisierung und automatisches Feedback neuer, abgeleiteter Daten in die Infrastruktur beinhalten und all diese Funktionalität ohne Softwareinstallation beim Benutzer web-basiert bereitstellen.

Der Aufbau und nachhaltige Betrieb einer derartigen Daten-und Analyseinfrastruktur sollte als eine nationale Aufgabe im Rahmen der KI-Strategie der Bundesregierung angesehen werden, um die Rolle Deutschlands als starken Wirtschafts- und Wissenschaftsstandort in einer digitalisieren Welt nachhaltig zu sichern. Eine derartige Initiative könnte in Analogie zum Flugzeugbau als „IT-Airbus“ bezeichnet werden: Analog zum Flugzeugbau gilt es, den Rückstand gegenüber amerikanischen und asiatischen Unternehmen durch einen koordinierten und kraftvollen Anschub des Staates aufzuholen. Gleichzeitig spielen Daten als Produktionsfaktor von nationalem Interesse eine Sonderrolle, so dass der Betrieb einer derartigen Infrastruktur als hoheitliche Aufgabe für die notwendige Neutralität und Vertrauen in Verfügbarkeit und Sicherheit sinnvoll sein könnte, insbesondere, um breiten Einsatz und Nutzung derselben Infrastruktur in unterschiedlichen Branchen und durch potentiellen konkurrierende Unternehmen sicherzustellen. Die Realisierung wird konzertierte Beiträge aus nahezu allen Bereichen der Informatik in Forschung und Entwicklung erfordern und sollte ein kurz-, mittel- und langfristiges Ziel der Gesellschaft der Informatik sein. Im Rahmen der KI Strategie der Bundesregierung ist eine derartige Infrastruktur vorgesehen [13].

Referenzen

[1] Apache Flink, https://flink.apache.org/

[2] Rapidminer, https://rapidminer.com/

[3] OpenML, https://www.openml.org/

[4] Sloan Digital Sky Survey: http://www.sdss.org

[5] Tony Hey, Stewart Tansley, Kristin Tolle, The Fourth Paradigm, https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

[6] Big Data Open Storage Network, https://hub.jhu.edu/2018/06/07/big-data-open-storage-network-alex-szalay/

[7] Rat für Informationsinfrastrukturen, http://www.rfii.de/de/themen/

[8] Rat für Wirtschafts- und Sozialdaten, https://www.ratswd.de/forschungsdaten/fdi-ausschuss

[9] NFDI Arbeitsgruppe, https://www.akademienunion.de/arbeitsgruppen/ehumanities/nfdi-arbeitsgruppe/

[10] International Data Space, https://www.fraunhofer.de/de/forschung/fraunhofer-initiativen/industrial-data-space.html

[11] Scads, https://www.scads.de/de/

[12] BBDC, http://www.bbdc.berlin/home/

[13] KI Strategie, https://www.ki-strategie-deutschland.de/home.html

Über Prof. Volker Markl

Prof. Dr. Volker Markl leitet das Fachgebiet Datenbanksysteme und Informationsmanagement an der Technischen Universität Berlin und ist Chief Scientist und Leiter der Forschungsgruppe „Intelligente Analyse von Massendaten“ am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Berlin. Er ist Direktor des vom Bundesministerium für Forschung und Bildung eingerichteten

Kompetenzzentrums „Berlin Big Data Center“ (BBDC) zum Umgang mit großen Datenmengen, Co-Direktor des Kompetenzzentrums „Berliner Zentrum für maschinelles Lernen“ (BZML) und Direktor des Smart Data Forums des Bundesministeriums für Wirtschaft und Energie.

Prof. Dr. Markl war vor seiner Tätigkeit in Berlin neun Jahre als Forscher und Innovator bei IBM im Silicon Valley tätig. Parallel zu seiner wissenschaftlichen Laufbahn hat Prof. Dr. Volker Markl mehrere Startups begleitet bzw. mitgegründet, unter anderem Parstream (heute CISCO) und dataArtisans. Die dataArtisans sind aus dem Stratosphere-Projekt der TU Berlin hervorgegangen und haben die Datenstromverarbeitungstechnologie Apache Flink erfolgreich weiterentwickelt und kommerzialisiert. 2014 wurde Prof. Dr. Markl als ein der führenden „Digitalen Köpfe“ Deutschlands von der Gesellschaft für Informatik ausgezeichnet.