

BIG DATA IN ÖFFENTLICHER IT



Ansprechpartner

Jens Fromm
Leiter Kompetenzzentrum
Öffentliche IT
Tel. +49 (0)30 3463-7173
Fax +49 (0)30 3463-99-7173
jens.fromm@fokus.fraunhofer.de

Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin

www.fokus.fraunhofer.de
www.oeffentliche-it.de



Grundlagen

Big Data beschäftigt sich mit der intelligenten Auswertung großer Mengen digitaler Daten. Mit intelligenter Auswertung ist hier gemeint, dass Daten aus unterschiedlichsten Quellen gesammelt, aufbereitet, zusammengeführt, analysiert und zu Informationen veredelt werden. Dies erhöht das Wissen der Auswerter signifikant und ermöglicht es, strategische Entscheidungen auf breiter Wissensgrundlage zu treffen. Bei großen Mengen kann es sich um das in heutigen Datenzentren typischerweise vorhandene Datenvolumen von Tera- und Petabytes, aber auch um das tausend- oder millionenfache Datenvolumen von Exa- und Zettabytes handeln, das die Gesamtmenge der heute überwiegend in sozialen Medien im Internet vorhandenen digitalen Daten beschreibt. Daten, das sind vermehrt die von uns wissentlich und unwissentlich bei der Nutzung sozialer Netzwerke und beim Surfen im Internet hinterlassenen Spuren unserer Netzaktivitäten oder die bei der Nutzung

mobiler Endgeräte erfassten Bewegungsdaten, das sind auch die Daten eingebetteter Systeme und Sensoren aus Bereichen wie Medizin, Logistik, Verkehr und Energie, das sind Produktions- und Maschinendaten oder auch die Ergebnisse klassischer kommerzieller, technischer und wissenschaftlicher Datenverarbeitung.

Technologien

Big-Data-Technologien unterstützen den Umgang mit großen Datenmengen. Dazu gehören die Speicherung der Daten, das Filtern und logische Zusammenführen von Daten aus verschiedenen Quellen, die Bereitstellung von Verfahren zur Analyse der Daten, die Möglichkeit zur Formulierung dedizierter Anfragen und letztendlich die verständliche Darstellung der Antworten auf diese Anfragen. Wo befindet sich hier aber das Neue gegenüber klassischen Verfahren und Prozessen? Aus dem skizzierten Spektrum der Datenvolumen über mehrere Zeh-



Abbildung: Typische Phasen von Datenauswertungen

nerpotenzen lässt sich einerseits ableiten, dass der Übergang von klassischen Analyseverfahren zu Big-Data-Konzepten ein fließender ist und damit eine scharfe Trennung der Konzepte unmöglich ist. Andererseits ergeben sich aus anderen charakteristischen Attributen weitere Unterscheidungskriterien. So können beispielsweise Sensoren einen kontinuierlichen Datenfluss liefern, der zeitnah zur Optimierung von Prozessen ausgewertet werden soll. Im Allgemeinen kann man davon ausgehen, dass Big Data zumindest durch ein hohes Volumen oder aber durch eine hohe Änderungsrate charakterisiert ist. Die Beschaffenheit der betrachteten Daten stellt eine weitere Herausforderung dar. Sogenannte unstrukturierte Daten wie Texte, Bilder oder Videos können syntaktisch und semantisch analysiert und zur Beantwortung von Fragen mit herangezogen werden. Die nutzbaren Daten können und müssen durch Vorverarbeitung und Filterung von »Datenmüll« getrennt und auf ihre Qualität zur Beantwortung der gestellten Fragen analysiert werden. Um diese Aufgaben in angemessener Zeit durchzuführen, spielen moderne IKT-Infrastrukturen wie schnelle Kommunikationsnetze, dedizierte Speicher-technologien und durch Cloud-Computing bereitgestellte Speicherkapazitäten und Rechenleistungen eine wesentliche Rolle.

Potenziale

Sind wir heute in der Lage, die Potenziale der großen Datenmengen überhaupt zu erkennen und die Fragen zu stellen, die die Analyse vorliegender Daten sinnvoll beantworten könnte? Big-Data-Analysen liefern

mathematisch untermauerte Aussagen über Gemeinsamkeiten, Anomalien, Zusammenhänge, Ursache-Wirkung-Beziehungen, Statistiken, Prognosen und Optimierungspotenziale und -strategien. Wir müssen lernen und lehren, diese Potenziale zu identifizieren und ihre Auswirkungen für Entscheidungsträger aus Politik und Wirtschaft, aber auch für die Zivilgesellschaft richtig einzuschätzen. Auf diesem Weg dürfen wir uns von den technischen Möglichkeiten jedoch nicht blenden lassen, denn wie bei jeder Entwicklung werden auch bei Big Data die positiven Nutzungsszenarien und der potenzielle Missbrauch kontrovers diskutiert.

Herausforderungen

Big Data lebt von der Extraktion und Integration von Informationen aus verschiedenen Quellen. In vielen Fällen werden öffentlich verfügbare Daten aus dem Internet mit privaten Daten gemeinsam ausgewertet, um Prognosen und Strategien zu entwickeln oder um eigene Vorhersagen und Modelle zu bestätigen, zu widerlegen oder hinsichtlich optimaler Lösungen miteinander zu vergleichen. Dem Zusammenführen personenbezogener und personenbeziehbarer Daten aus verschiedenen Quellen sind jedoch auch enge rechtliche und ethische Grenzen gesetzt. Die Grundidee der Sammlung und Auswertung von personenbezogenen Daten aus verschiedensten Quellen widerspricht fundamental den Prinzipien der Zweckbindung bei der Erhebung und Speicherung von Daten und der Datensparsamkeit. Bundesdatenschutzgesetz, Telemediengesetz aber auch die Europäische Datenschutz-

richtlinie definieren einen engen Rahmen für den Einsatz von Big Data im öffentlichen Sektor. Der Bürger muss als Individuum durch Anonymisierung und Pseudonymisierung seiner Daten davor geschützt werden, zum gläsernen Menschen zu werden. Auch seine Identität, Privatsphäre und Reputation müssen gegen Manipulation durch Dritte geschützt werden.

Chancen

Bei allen unbestreitbaren Risiken dürfen die mit einem verantwortlichen Einsatz von Big Data verbundenen Chancen nicht übersehen werden. Um diese Potenziale zu kommunizieren, ist die Entwicklung und Demonstration von sinnvollen Anwendungen, d.h. erst durch Big Data umsetzbarer Szenarien, ein unverzichtbarer Schritt. Überzeugende Anwendungsbeispiele wichtiger Handlungsfelder aus dem öffentlichen Raum, wie z.B. Open Government, Smart Energy oder Mobilität sowie intelligente Verkehrs- und Transportsysteme, sind für eine erfolgreiche Etablierung von Big-Data-Konzepten erforderlich.

Beispiele aus Anwendungsfeldern:

- Transparenz mit Open Data
- Trenderkennung mit sozialen Medien
- Optimierung mit dynamischen Daten
- Steuerung mit Sensordaten
- Sicherheit mit verteilten Messdaten
- Optimierung mit Transaktionsdaten