Big Data – Chancen und Herausforderungen

Von Roberto V. Zicari

Jeden Tag werden 2,5 Trillionen Bytes an Daten generiert. Diese enorme Menge an Daten wird beispielsweise durch digitale Bilder, Videos, Beiträge in den sozialen Medien, intelligente Sensoren, Einzelhandels- und Finanztransaktionen und GPS-Signale von Handys erzeugt. Das ist Big Data.
Es besteht kein Zweifel daran, dass Big Data und das, was wir damit tun, das Potential hat, ein signifikanter Treiber für Innovationen und Wertschöpfung zu werden.

Definition und Chancen

Was ist Big Data? Für die Beantwortung der Frage habe ich eine Definition des McKinsey Global Institute (MGI) gewählt: „Big Data bezieht sich auf Datenmengen, deren Größe über die Fähigkeiten typischer Datenbankanwendungen hinausgeht, diese zu erfassen, zu speichern und zu analysieren“. Wo finden wir Big Data? Daten im Allgemeinen, und vor allem *Massendaten*, so ein älterer Begriff für Big Data, sind ein wichtiger Produktionsfaktor in allen Industrien und Business-Prozessen. MGI schätzt, dass etwa 7 Exabyte an neuen Daten von Unternehmen im Jahr 2010 gespeichert wurden. Interessanterweise wird 50 Prozent des IP-Datenverkehrs nicht von Menschen verursacht, sondern von Maschinen. Zudem wird die *Maschine-zu-Maschine-Kommunikation* (M2M) zunehmend bedeutsamer.

Doch worin liegt letztlich der Nutzen von Big Data? Big Data an sich erzeugt noch keinen Mehrwert. Erst durch die Analyse der gigantischen Datenmenge kann ein Mehrwert erzeugt werden: die Erzeugung von Transparenz; die Identifizierung von Bedürfnissen, Verdeutlichung von Veränderungen, Verbesserung von Leistungen; die Segmentierung von Kunden; und schließlich die Unterstützung der menschlichen Entscheidungsfindung mit automatisierten Algorithmen, innovativen und neuen Geschäftsideen, mit Produkten und Services.

Die Fähigkeit, interaktive Datenexploration mit Datenanalyse und -visualisierung zu kombinieren, ist in Bezug auf Big Data besonders wichtig, da dies zu neuen Erkenntnissen führt, die ansonsten unentdeckt blieben. Durch die Möglichkeit, große Mengen komplexer Daten aus unterschiedlichen Quellen interaktiv zu explorieren, können Organisationen neue Erkenntnisse über ihre Produkte, Kunden und Services gewinnen.

Das Konzept „Big Data Suche“ impliziert dabei, dass die Art und Weise nicht festgelegt ist, um Erkenntnisse aus Big Data zu gewinnen. Die Vorgehensweise hängt stark vom jeweiligen Anwendungsfall ab. Häufig führt bereits die einfache interaktive Exploration von großen Mengen komplexer Daten aus verschiedenen Quellen zu neuen Erkenntnissen über Produkte, Kunden und Services.

Wo kommt Big Data zum Einsatz? Big Data kann einen industrieübergreifenden finanziellen Mehrwert generieren. Zu den Schlüsselbereichen zählen:

- das Gesundheitswesen – ein sehr heikler Bereich bezüglich des Datenschutzes,
- der Öffentliche Sektor – zum Beispiel öffentlich zugängliche Daten („Open Data“) in Europa,
- globale und persönliche Standortdaten – besonders relevant für mobile Geräte,
- der Einzelhandel – interessant für große Onlineportale wie eBay und Amazon,
- die Produktion,
- soziale Daten – sowohl persönliche als auch berufsbezogene Daten aus sozialen Netzwerken wie Facebook oder Twitter.

Es gibt unzählige Beispiele, in denen Big Data zur Anwendung kommt. Die Relevantesten sind Log Analytics (Analyse der Ereignisprotokolldatei eines Computersystems), Betrugserkennung, Social Media- und Sentiment-Analyse („Stimmungserkennung“ eines Textes), Risikomodellierung, Unternehmensführung und Energie-Management. Eines der Hauptprobleme bei der Nutzung von Big Data in Unternehmen ist derzeit der Mangel an Fachkräften mit dem notwendigen Wissen und den Fähigkeiten in den Bereichen Statistik, Machine Learning und Data Mining.

Foto: Fotolia, Rafał Olechowski

Drei große Herausforderungen

Es wurde bereits auf die Potenziale von Big Data für die Ökonomie eingegangen. Doch wie sieht die Realität heutzutage aus? Die künftigen Herausforderungen im Umgang mit Big Data können in drei Dimensionen eingeteilt werden: *Daten, Prozesse, Management*.

Daten
Die Hauptherausforderung ist der Umgang mit der großen Menge an Daten, also mit dem *Volumen*. Eine weitere Schwierigkeit besteht im Umgang mit unterschiedlichen Datentypen, -quellen und -formaten, also mit der *Vielfalt* und der *Kombination* von verschiedenen Datensätzen. Ebenso stellt sich die Frage nach der *Geschwindigkeit*: Wie kann auf die Informationsflut in angemessener Zeit reagiert werden? Darüber hinaus müssen wir uns auch mit den Aspekten des *Wahrheitsgehalts*, der *Datenqualität* und *-verfügbarkeit* auseinandersetzen: Wie können wir mit Unsicherheiten, fehlenden Werten und falschen Angaben umgehen? Wie „gut“ sind die Daten, gibt es überhaupt verfügbare Daten und wie gut ist die Stichprobe? Ein weiteres Feld betrifft das *Auffinden* von Daten: Wie lassen sich hochqualitative Daten in der gigantischen Menge von Daten im Web finden? Fragt man beispielsweise, ob in den Datensätzen bestimmte zugrunde liegende Annahmen getroffen werden, dann ist die *Qualität* der Datensätze und deren *Relevanz* für bestimmte Probleme angesprochen.

Auch der Aspekt der *Vollständigkeit* ist zu bedenken: Decken die Daten den gesamten Anwendungsfall ab? Was impliziert dies? Unter dem Schlagwort der *Anonymisierung* wäre wiederum die Frage zu stellen: Können wir aus Daten genug Informationen extrahieren, ohne die Identität von Personen aufzudecken oder den Datenschutz zu verletzen? Die Analyse von Big Data kann zu bemerkenswerten Erkenntnissen führen. Doch wir müssen gewährleisten, dass wir den Ergebnissen nicht blind vertrauen, sondern weiterhin Domänenexpertise und Sachverstand mit einfließen lassen, uns also gegen *Daten-Dogmatismus* schützen. Und schließlich geht es auch um Skalierbarkeit: Dies umfasst Techniken wie *Social Graph Analysis*. In größeren Graphen kommen aktuelle Technologien schnell an die Grenzen des Machbaren.

Die Kombination all dieser Probleme führt zu einer Vielzahl von Herausforderungen und Chancen, um, verglichen mit traditionellen Ansätzen, schnellere, bessere und günstigere Lösungen für *Big Data Analytics* zu finden.

Prozesse
Eine große Herausforderung in diesem Kontext ist die Frage, wie die Daten analysiert werden können. Die Auswahl des richtigen Analyse-Modells kann beträchtliche Zeit in Anspruch nehmen. Die Fähigkeit, verschiedene Modelle schnell zu testen, um das beste Modell zu finden, ist von besonderer Bedeutung. Die Herausforderungen in Bezug auf den Erkenntnisgewinn beinhalten:

- die Datenerfassung,
- das Angleichen von Daten aus verschiedenen Quellen (zum Beispiel um Duplikate zu identifizieren),
- die Transformation von Daten in eine für die Analyse geeignete Form,
- die Modellierung, entweder mathematisch oder durch Simulation,
- das Verständnis für den Output, Visualisierung und Teilen der (komplexen) Ergebnisse.

Management
Die größten Herausforderungen des Daten-Managements sind verbunden mit den Themen Datenschutz, Sicherheit, Verwaltung und mit ethischen Fragestellungen. Die größten Herausforderungen für das Management sind die Sicherstellung richtiger Datennutzung; das bedeutet, dass es die beabsichtigten Zwecke und die geltenden rechtlichen Bestimmungen sowie die Nachverfolgung, wie die Daten genutzt, umgewandelt und abgeleitet werden, festzuhalten gilt. Gleiches gilt für den Lebenszyklus der Daten. Viele Datenbanken enthalten sensible Daten, wie beispielsweise Personaldaten. Es gibt rechtliche und ethische Bedenken in Bezug auf den Zugang zu solchen Daten. Somit gilt es sicherzustellen, dass die Daten geschützt werden und der Zugang kontrolliert und zur Überprüfung protokolliert wird.

Fazit

Big Data ist mittlerweile kein inhaltsleeres Modewort mehr, sondern es ist der Schlüssel zu Innovation und hat ein enormes Potential für die Wertschöpfung. Es gibt eine Vielzahl an Möglichkeiten – zum Beispiel für das Gesundheitswesen – ortsbezogene Daten, den Handel, die Produktion oder gesellschaftliche Daten. Zudem gibt es eine Reihe von Herausforderungen, zum Beispiel in Bezug auf Datenvolumen, Datenqualität, Datenerfassung und Datenmanagement, ebenso wie Datenschutz, Sicherheit und Verwaltung. Ich möchte zum Abschluss dieses Artikels auf eine interessante Chance für Big Data hinweisen:

„As more data become less costly and technology breaks barriers to acquisition and analysis, the opportunity to deliver actionable information for civic purposed grows. This might be termed the ‚common good‘ challenge for big data.“ (Jake Porway, DataKind)

Weitere Informationen
Ringvorlesung zu Big Data, Internet of Things und Data Science, Frankfurt Big Data Lab, 2015;

Video-Mitschnitte der Vorlesungen:
➤ www.bigdata.uni-frankfurt.de/soft-skills-entrepreneurship-m-ssk-b-sos

Frankfurt Big Data Lab:
➤ www.bigdata.uni-frankfurt.de

Foto: privat

Prof. Roberto V. Zicari ist seit 1992 Professor für Datenbanken und Informationssysteme an der Goethe-Universität.

Zicari ist Herausgeber des Portals ODBMS.ORG (Operational Database Management Systems):
➤ www.odbms.org

Überblick

Aktuell 2
Forschung 6
Reportage 12
International 14
Kultur 15
Campus 16
Impressum 23
Bücher 24
Bibliothek 25
Freunde 26
Studium 27
Menschen 30
Termine 31

Im nächsten UniReport ...
... wird nun endlich der naturwissenschaftliche Campus auf dem Riedberg porträtiert.
Ausgabe 6/2015 erscheint am 3.12.2015, Redaktionsschluss ist am 10.11.2015