Big Data

Begriffsdefinition

Es existiert keine eindeutig festgelegte Definition für den Begriff „**Big Data**“. Es setzt sich aus den englischen Worten *big* für „groß“ und *data* für „Daten“ zusammen. Es liegt daher nahe, dass man Big Data mit dem Ausmaß und Umfang an Daten assoziiert. Gemeinhin versteht man unter Big Data große Mengen an strukturierten, semi-strukturierten oder unstrukturierten Daten, die so enorm und komplex sind, dass der Umgang mit ihnen unter Nutzung traditioneller Techniken und Methoden der Datenverarbeitung nicht mehr zu bewältigen ist.

Die Formulierung „Big Data“ bezieht sich jedoch nicht ausschließlich auf die Größenordnung der Datensätze, sondern wird oft auch synonym verwendet für die Speicherung, Verarbeitung und Analyse dieser Daten unterschiedlichster Herkunft und Formate zur Informationsgewinnung unter Zuhilfenahme neuer Technologien, Praktiken und Anwendungen.

Suchtrends und Definitionen großer Technologiekonzerne zu Big Data untersuchend, versucht eine Studie der University of St. Andrews, eine möglichst allgemeingültige Beschreibung für “Big Data” zu definieren:

„Big Data ist ein Begriff, der die Speicherung und Analyse großer und oder komplexer Datensätze unter Verwendung einer Reihe von Techniken beschreibt, die unter anderem NoSQL, MapReduce und Machine Learning beinhalten“.

(Original: *„Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning“* (Ward/Barker 2013: S. 2).

Hintergrund

Vom Anbeginn der Zeitrechnung bis zum Jahre 2003 wurden etwa fünf Milliarden Gigabyte an Daten erschaffen. Im Jahre 2011 entstand diese Menge bereits in nur zwei Tagen und ab dem Jahre 2013 schon alle zehn Minuten (Heuer 2013: S. 6). Diese Datenmenge verdoppelt sich alle zwei Jahre. Bis zum Jahre 2020 soll das Volumen der der jährlich erzeugten und kopierten Daten auf 44 Zettabyte ansteigen (Turner et al. 2014: S. 1). Diese Explosion an Datenmassen ist nicht zuletzt auch dem menschlichen Bedürfnis nach Unterhaltung, Verbindung und Kommunikation geschuldet. Wurde früher der Großteil der Daten durch Transaktionssysteme erzeugt, so kommt heute die Mehrheit der Datenflut aus sozialen Netzwerken. Das Internet hat sich seit seiner Erschaffung von einem Netzwerk für Forschungszwecke zu einem weltweiten Kommunikationsnetzwerk weiterentwickelt. Zudem finden durch die zunehmende Verbreitung von Computern nun immer mehr Menschen Zugang zum Internet. Besonders die Entwicklung und die Beliebtheit internetfähiger Mobilgeräte wie Smartphones und Tablets haben in den letzten Jahren das Wachstum weiter vorangetrieben und tragen so zu einer ständigen Vernetzung der Menschen und damit auch zur Erzeugung neuer Daten bei.

Zu Big Data zählen aber nicht nur Kommunikationsdaten von Personen untereinander, sondern auch von Mensch-Maschine- oder Maschinen-Maschinen-Interaktionen und vermehrt auch von eingebetteten Computern in „intelligenten“ Geräten, die den Alltag erleichtern sollen und dazu

diverse Daten in einem „Internet der Dinge“ versenden und empfangen. Dazu kommen auch Daten wie automatisch generierte Logfiles, Sensordaten, Daten von Überwachungsnetzwerken, Wetterstationen, RFID-Chips, GPS-Geräten, Ampelanlagen oder Verkehrsregelungssystemen. Die zunehmende Technisierung und Vernetzung der Gesellschaft geht einher mit einer ununterbrochen anhaltenden Entstehung neuer Daten und tragen so zu einer stetig wachsenden Datenflut bei. Dieses Phänomen stellt die Datenverarbeitung –und Analyse vor ganz neue Herausforderungen. Oft werden diese mit dem Drei-V-Modell beschrieben (vgl. Laney 2001):

1. **Volume**: Meint die Größenordnung der Datenmengen.
2. **Variety**: Beschreibt die Verschiedenartigkeit der Daten.
3. **Velocity**: Die Geschwindigkeit, in der neue Daten anfallen und verarbeitet werden.

IBM erweitert dieses Modell mit „Veracity“ (Wahrhaftigkeit) um eine zusätzliche Dimension (IBM o. J.):

- **Veracity**: Bezieht sich auf die Qualität und Zuverlässigkeit der Daten.

Volume

Im Jahre 2014 wurden jede Minute rund 72 Stunden Videomaterial auf YouTube hochgeladen, 277.000 Tweets getwittert, 4 Millionen Google-Suchanfragen aufgegeben und 204 Millionen E-Mails versendet (DOMO 2014). Die Dimensionen, in denen Daten anfallen und erhoben werden, führen zu zwei Problemen: Zum einen bedeutet ein zu großes Datenvolumen ein Speicherproblem. Jedoch ist Speicher im Laufe der Zeit und im Zuge des technischen Fortschritts immer günstiger geworden. Das erlaubt es, massenweise Daten zu speichern. Für Unternehmen bietet sich die Option, die Daten im einem eigenem Datenzentrum im Haus zu sichern oder sie in einem externen Cloudspeicher auszulagern. Speziell für den schnellen und effizienten Umgang mit großem Datenvolumen wurden Open Source Plattformen wie Hadoop oder diverse NoSQL-Datenbanksysteme entwickelt. Cloudspeicher in Kombination mit Open Source Werkzeugen erlauben es nun jedermann, kostengünstig mit Daten im großen Stil zu arbeiten. Der größte Teil des weltweiten Datenvolumens ist jedoch flüchtiger Natur (Musik- oder Filmstreaming, Online-TV, Sensorsignale, etc.), d.h. die Daten werden nur kurzfristig oder gar nicht gespeichert, sondern gleich verarbeitet oder gelöscht. Nur etwa 33% hätten 2013 gespeichert werden können und 2020 wären es weniger als 15%, da die Speicherentwicklung mit der rasanten Datenwachstumsrate nicht mithalten kann (Turner et al. 2014: S. 3). Neben etwaiger Speicherprobleme führen große Mengen an Daten jedoch zusätzlich zu einem Datenanalyse-Problem. Für die Informationsgewinnung ist nicht unbedingt die Datenmenge entscheidend, da meist nur ein geringer Anteil der Daten überhaupt relevant ist, wichtiger ist, auch die richtigen Daten zu haben. Im Jahre 2013 waren nur etwa 5% aller Daten tatsächlich von Interesse (Turner et al. 2014: S. 2). Um Kosten für Speicher und Aufwand für Analysen einzudämmen, stellt sich also die Frage, welche Daten überhaupt erhoben und gespeichert werden sollen.

Variety

Ein besonderes Merkmal von Big Data ist sicherlich die Vielzahl an unterschiedlichen Daten, die anfallen können. Traditionelle Relationale Datenbankmanagementsysteme (RDBMS) sind für die Arbeit mit Daten in strukturierter Form ausgelegt. Sie speichern Daten nach einem festen Schema. Dazu werden die Inhalte in Beziehungen zueinander gestellt und Daten dann in strukturierten Tabellen, den Relationen, gespeichert. Mit der zunehmenden Zahl der verfügbaren Datenquellen kann aber keine geordnete Struktur und Homogenität der Daten mehr sichergestellt werden. Massenhafte Datenerhebung aus unterschiedlichsten Quellen führen dazu, dass auch unterschiedlichste Datenformate anfallen, von klassisch strukturierten, wie etwa Kundenstammdaten, die eine feste

Form aufweisen und problemlos verarbeitet werden können, über semi-strukturierte Daten, wie beispielsweise E-Mails, die Sender- und Empfängerinformationen in festgelegter Form aufweisen, deren Inhalt aber keinen festen Vorgaben folgt und zudem verschiedene Anhänge haben kann, bis hin zu völlig unstrukturierten Daten, in Form von Beiträgen auf Sozialen Netzwerken oder Videos, Bilder, Text, Audio oder gar Aufzeichnungen von Fitness-Apps. Bei Big Data werden nun aber sämtliche Daten, ungeachtet ihrer Struktur, zusammengefasst und analysiert. Dabei können die zusammengefassten Daten ihrem Ursprung entsprechend in Kategorien unterteilt werden, wobei die Daten entweder durch Mensch-Mensch-, Mensch-Maschinen- oder Maschinen-Maschinen-Kommunikation entstanden sein können. (Klein et al. 2013)

Velocity

Dem Volumen ähnlich, verhält es sich mit der Geschwindigkeit. Ein exponentielles Datenwachstum bedeutet gleichermaßen eine erhöhte Datenwachstumsrate, es entstehen mehr Daten in kürzerer Zeit. Dabei gilt es zwei Faktoren zu betrachten: Zum einen die Geschwindigkeit, in der neue Daten anfallen und zum anderen wie zeitnah diese verarbeitet werden können. Etwa 1,7 Megabyte an Daten wurden vergangenen Jahres (2014) pro Minute und pro Person erzeugt (Turner et al. 2014: S. 3). Zusätzlich verdoppelt sich das Datenwachstum alle zwei Jahre. Das lässt erkennen, dass eine Adaptierung der Datenverarbeitung notwendig wird, um mit dem Wachstum mithalten und effiziente Wertschöpfung aus den Daten betreiben zu können. Während Daten früher hauptsächlich durch Transaktionen entstanden sind und intern gespeichert wurden und deren Verarbeitung trotz längerer Wartezeit von Wert waren, fallen heute Daten in Echtzeit an, die nur wertvolle Informationen liefern können, wenn sie entsprechend schnell verarbeitet werden. Der Trend geht hin zu ad hoc Datenauswertungen. Ein Extrembeispiel für die Geschwindigkeit, in denen Daten entstehen und verarbeitet werden müssen, ist der Large Hadron Collider des CERN. Dort fallen etwa 600 Terabyte an Sensordaten pro Sekunde an. Doch nur ein Bruchteil davon ist wirklich relevant für die Forschung und muss gespeichert werden. In zwei Analyseverfahren wird diese enorme Datenmenge auf rund 1050 Megabyte (pro Sekunde) reduziert. Der Rest der Daten wird gelöscht (CERN 2012).

Veracity

Steht für die Verunsicherung über Datenqualität und/oder deren Korrektheit. Es sollte bedacht werden, dass unter Umständen nicht alle Daten immer so schnell aufbereitet oder bereinigt werden können, wie sie eintreffen. Das bedeutet eventuell, dass fehlerhafte Datensätze mitverarbeitet werden und zu ungenauen oder falschen Ergebnissen führen können. Zudem gibt es keine Garantie dafür, dass grundsätzlich alle gesammelten Daten und Informationen auch korrekt sind. Absichtliche Fehlinformationen oder Falschangaben sind keine Seltenheit, wenn etwa Nutzer online ihre Privatsphäre schützen möchten, indem sie in Formularen Falschangaben machen. Gerade auch auf Social Media Plattformen ist vermehrt mit übertriebener Selbstdarstellung und unrichtigen Angaben zu rechnen. Mangelhafte Datenqualität kostet den US-Wirtschaft jährlich über 3 Billionen Dollar (IBM o. J.). Dieser Punkt kann also neben dem Wahrheitsgehalt durchaus auch für den Wert der Daten stehen.

Entwicklungen

Klassische relationale Datenbankmanagementsysteme sind seit den 70er Jahren im Einsatz und wurden über die Jahrzehnte immer weiterentwickelt und optimiert. Sie speichern Daten nach einem festen Schema in Tabellen, den Relationen, ab und erlauben Zugriff und komplexe

Verarbeitungsanweisungen über die Abfragespreche SQL. Diese Systeme sind bestens für Anwendungsfälle geeignet, bei denen die Daten in geordneter Form und Struktur vorliegen. Um nun aber den wachsenden Big Data Herausforderungen begegnen zu können, also dem Anfallen von enormen Datenmengen in kürzester Zeit, dem Zugriff durch bis zu mehrere Millionen Nutzer, der Forderung nach Echtzeitanalysen oder dem Umgang mit unstrukturierten Daten und schemafreien Dokumenten, haben einige Entwicklungen stattgefunden, welche die klassischen relationalen Datenbanken, wenn nicht ersetzen, so doch sinnvoll unterstützen und ergänzen können.

Siehe dazu:

- In-Memory
- NoSQL
    - Key-Value-Datenbanken
    - Dokumentenorientierte Datenbanken
    - Spaltenorientierte Datenbanken
    - Graphdatenbanken
    - MapReduce
    - Hadoop
- NewSQL

From:
https://gpm.wi-wiki.de/ - **Wirtschaftsinformatik Wiki - Kewee**

Permanent link:
https://gpm.wi-wiki.de/doku.php?id=bigdata:big_data

Last update: **2015/10/29 20:15**