1 Big Data und die sechs V-Herausforderungen

Im Zusammenhang mit Big-Data-Definitionen werden drei bis vier Herausforderungen beschrieben, die jeweils mit V beginnen. In der ursprünglichen Definition zu Begin des 21. Jahrhunderts wurden nur drei Begriffe genannt: Volumen, Variety und Velocity. Volumen steht dabei für die Größe der Datenmenge bei unstrukturierten Daten: Die notwendige Datenanalyse kann nicht mehr mit herkömmlichen Mitteln bewältigt werden. Variety steht für die Vielfalt der Datenformate und Datenquellen, die durch komplexe Verarbeitungsprozesse im Unternehmen nicht mehr unter einen Hut gebracht werden können. Velocity steht für die zeitgerechte Verarbeitung der Datenmengen, damit schnell Resultate für Entscheidungsprozesse vorliegen. Recht früh konnte man aber in der Big-Data-Diskussion feststellen, dass nur dann gute Resultate mit Big Data erzielt werden können, wenn die zugrundeliegende Datenqualität ausreichend gesichert ist. Der vierte V-Begriff (Veracity) ist damit integraler Bestand von Big-Data-Definitionen geworden. Was steckt aus Unternehmersicht hinter den vier V-Begriffen und wie müssen Big-Data-Konzepte mit diesen Herausforderungen im Unternehmensinteresse klar kommen?

1.1 Volumen: (Datenvolumen)

Die Datenmenge kann nicht mehr mit herkömmlichen Mitteln bewältigt werden

Bisher war es so, dass große Datenmengen in isolierten Datensilos für sich getrennt betrachtet wurden. Jeder Bereich hatte seine eigenständigen Aufgaben und Datenbanken hatten den Zweck, diesen Bereich umfassend für die bereichsspezifischen Zwecke abzubilden. So wurden Marketinginformationen für das Marketing aufbereitet, Zahlungsinformationen wurden für das Rechnungswesen genutzt. Doch inzwischen hat man erkannt, dass diese Trennung der Bereiche nicht angemessen ist, um zukunftsfähig zu bleiben. Relevante Informationen müssen aus verschiedenen Bereichen zusammengeführt werden, damit der Erfolg für das Unternehmen umfassend umgesetzt werden kann.

1.2 Variety: (Vielfalt)

Die Vielfalt von Datenquellen und Datenformaten erfordern eine andere Datenanalyse

Allerdings liegen in den unterschiedlichen Bereichen die Daten sehr unterschiedlich vor. Benutzerinteraktionen auf Webseiten können nicht sehr einfach den Kunden oder Interessenten zugeordnet werden. Bei Kommunikation mit Kunden werden Textinformationen benutzt, die nicht zu vorgegebenen Datenfelder passen. Es besteht die Gefahr, dass isolierte Datenbestände nicht so genutzt werden können, wie es für zukunftsorientiertes Handeln notwendig ist. Die Vielfalt der Datenformate kann aber nicht einfach reduziert werden, weil dadurch die Arbeit in den einzelnen Bereichen unzulässig eingeschränkt würde. (z.B. unstrukturierte Daten, Bilder, Texte, Videos)

1.3 Velocity: (Geschwindigkeit)

Die zeitgerechte Verarbeitung der Daten muss gesichert werden

Ein weiteres großes Problem ist der Zeitaufwand, wenn unterschiedliche Datenbestände zusammengeführt werden müssen.(Geschwindigkeit) Glücklicherweise sind die Kosten für Arbeitsspeicher inzwischen so niedrig, dass es machbar wird, auch sehr große Datenmengen gleichzeitig in den Zugriff zu nehmen. Man spricht in diesem Zusammenhang von In-Memory-Datenverarbeitung: Unterschiedliche Datenbanken werden gleichzeitig in den Speicher geladen und intelligente Computerprogramme greifen ein, um Zusammenhänge zu finden, an die bisher noch nicht gedacht werden konnte.

1.4 Veracity: (Wahrhaftigkeit)

Die Datenqualität entscheidet über den Erfolg von Big Data

Doch das Hauptproblem für Big-Data-Auswertungsprozesse ist, dass unstrukturierte Daten nicht unbedingt zutreffende Daten über die zugrundeliegenden Prozesse darstellen. Daten fallen in allen Geschäftsbereichen laufend an und werden je nach Anwendungsfall sehr unterschiedlich weiterverarbeitet. So werden Rückmeldungen von Interessenten gelegentlich dokumentiert, aber in nicht ausreichend differenzierten Datenbeständen, die später Schwierigkeiten bei der Auswertung machen. Unternehmen sollten daran arbeiten, die Vielzahl der Datenbestände hinsichtlich ihrer Relevanz zu bewerten. Handelt es sich um Meinungsäußerungen oder Kritik an Produkten oder Vorgehensweisen, dann sollte sichergestellt werden, dass diese Aussagen repräsentativ erfasst werden, denn sonst entsteht ein schiefes Bild. Viele Daten garantieren keine gute Auswertungsperspektive, wenn es eher zufällig ist, dass diese Daten so umgeformt werden, dass sie ausreichend genau und spezifisch die zugrundeliegenden Abläufe darstellen. (z.B. Kunden-/Produktbewertungen --> Speicherung in einheitlicher Bewertungsskala anstelle von Fließtext)

Mittlerweile sind noch **2 weitere wichtige V's** hinzugekommen, so dass man heute von 6 V spricht.

In der Literatur wird teilweise auch von 8 V, 10 V oder sogar 42 V gesprochen. Hier beschränken wir uns hier auf 6 V's. (https://datadrivencompany.de/big-data-definition-merkmale-technologien/#8_Vs_10_Vs_42_Vs)

1.5 Value (Wert)

Die Auswirkung auf das Business

Das fünfte V ist eigentlich nicht erklärungsbedürtftig: Es steht für den Wert, sprich die Verwertbarkeit der mit Big Data erschlossenen Daten. Die Szenarien dafür können je nach Branche ganz unterschiedlich sein. So lassen sich mit Big Data etwa Produktionsprozesse optimieren, neue Zielgruppen erschließen oder ganz neue Produkte entwickeln. Die Einsatzfelder von Big Data sind, ähnlich denen der Elektrizität, praktisch unbegrenzt.

1.6 Variability (Variabilität)

Die (mögliche) Mehrdeutigkeit der Daten

Das letzte V bezieht sich auf Variability. Variabilität bedeutet, dass manche Datensätze weniger konsistent als herkömmliche Transaktionsdaten sind und möglicherweise mehrere Bedeutungen haben oder von einer Datenquelle zur anderen unterschiedlich formatiert sind. Das sind Faktoren, die die Verarbeitung und Analyse der Daten erschweren.

YouTube-Video in englisch: https://www.youtube.com/watch?v=QbSFgdbBLkY

Quellen:
https://www.bde-gmbh.de/blog/big-data/die-herausforderung-der-4-vs/
https://www.micromata.de/