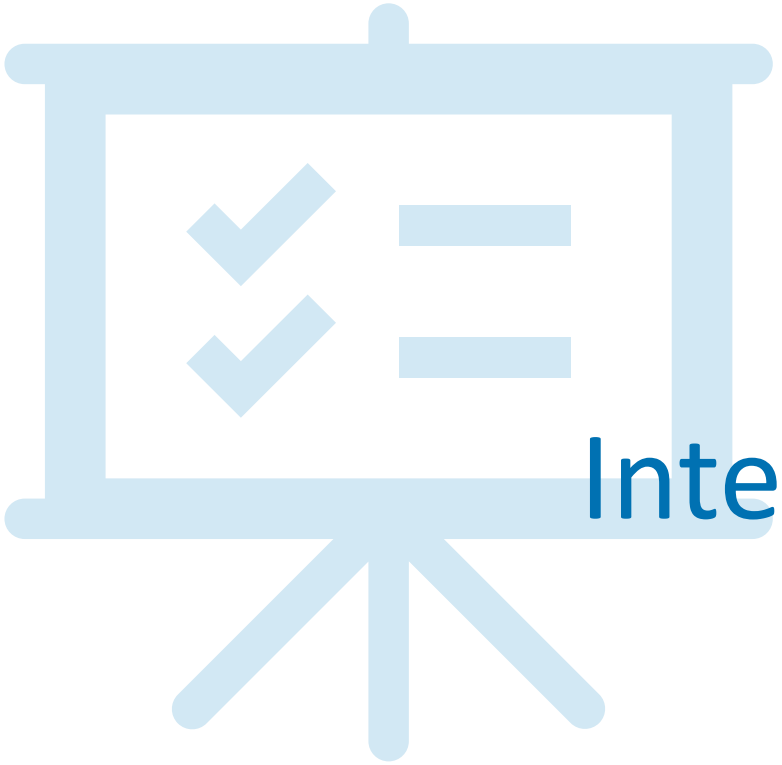


Netzwerke und Internettechnologien 2

Big Data

Netzwerke und Internettechnologien 2

Lernziele

1

Big Data

Grundlagen

2

ETL

Big Data

Bedeutung

- Das Auffinden von Mustern in den täglich anfallenden riesigen Datenmengen zu finden, bringt dem Nutzer enorme Vorteile für
  - die Gewinnung wissenschaftlicher Erkenntnisse.
  - die Festsetzung von Preisen.
  - das Generieren von Kaufempfehlungen.
  - die Aufdeckung verdächtiger Aktivitäten.
- Viele Unternehmen, Wirtschaftszweige und auch Behörden setzen jetzt auf „Big Data“.
- Aber was ist eigentlich „Big Data“?

Big Data

Bedeutung

- Big Data ist ein allgemeiner Begriff, für die Beschreibung sehr großer Mengen unstrukturierter und semi-strukturierter Daten, die Unternehmen ununterbrochen produzieren.
- Big Data bezieht sich nicht auf eine bestimmte Datenmenge, wird aber als Synonym für Peta- und Exabyte an Daten genutzt.
- Diese Daten in einer relationale Datenbank zu analysieren ist teuer und aufwändig.
- Der im Internet und in den Unternehmen verfügbare Datenberg (Big Data) wird immer größer, unübersichtlicher und lässt sich nur schwer verarbeiten.
- Tools und Programme mit neuesten Technologien erstellt, sollen das Problem lösen (Beispiele sind Apache Spark oder Hadoop).

Big Data

Dimensionen

- Big Data bezieht sich in der Definition auf 4 Dimensionen:

    1. Data Volume (Volumen)
    2. Data Velocity (Geschwindigkeit)
    3. Data Variety (Vielfalt)

- Erweitert wird diese Definition um die Dimensionen
    5. Data Veracity (Echtheit)
    6. Data Value (Mehrwert)

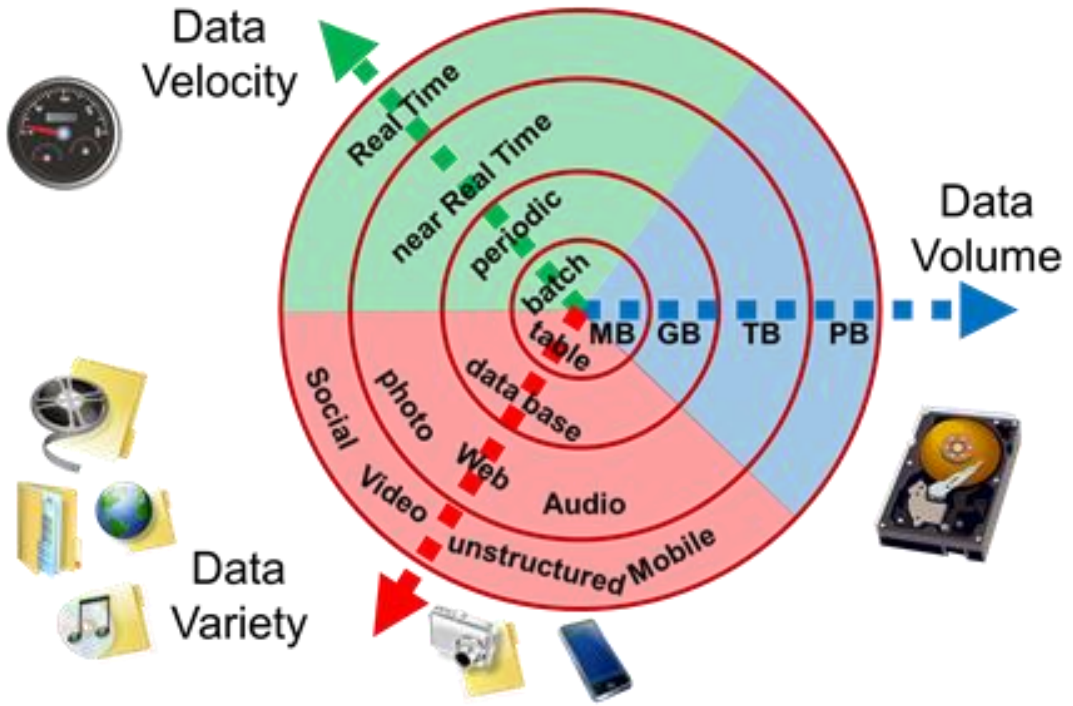
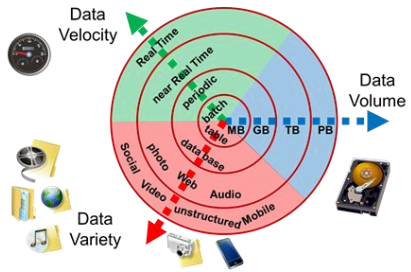


Abbildung 1: Big Data (Ender005 (https://commons.wikimedia.org/wiki/File:Big_Data.png))

Big Data

Dimensionen

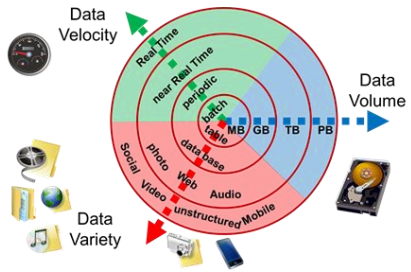


Data Volume

- Bei Big Data geht es darum große Mengen unstrukturierter Daten geringer Dichte verarbeiten, z.B. Trafficdaten auf einer Webseite, einer mobilen Anwendung oder sensoraktivierte Geräte, wie beim Internet der Dinge oder Themen der Industrie 4.0…
- Diese enormen Datenmengen stellen für herkömmliche Datenbanksysteme eine große Herausforderung dar, müssen aber bewältigt werden.
- Die Datenmengen werden auch in Zukunft kontinuierlich ansteigen, nach Statista 2017 bis 2025 auf 163 Zettabyte jährlich.

Big Data

Dimensionen

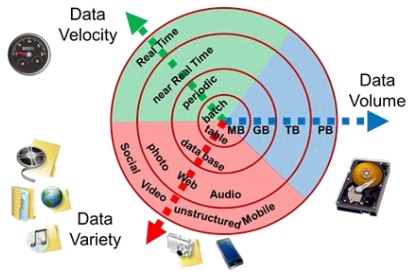


Data Velocity

- Für die Geschwindigkeit gibt es zwei Sichtweisen:
    1. Die enorme Erzeugungsrate von Daten in den verschiedensten Anwendungsfeldern.
    2. Die rasant wachsende Datenmenge muß, für eine schnelle Reaktion, auch zeitnah weiterverarbeitet werden.
- Software oder Geräte, die internetfähige intelligente Daten sammeln, liefern Daten in Echtzeit und erfordern sehr schnelle Echtzeitauswertung und Echtzeitaktionen.

Big Data

Dimensionen

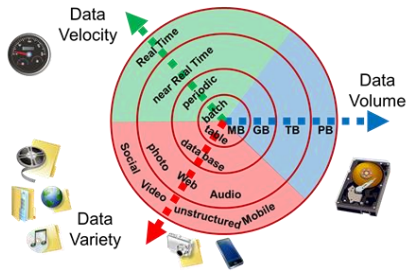


Data Variety

- Bezieht sich auf die vielen neuen Arten von un- und halbstrukturierte Datentypen, die zur Verfügung stehen.
- Um die Bedeutung dieser Datentypen, wie Text, Audio und Video zu verstehen und Metadaten für die Verwaltung und Analyse generieren zu können, ist zusätzlicher Aufwand erforderlich.
- Mit den herkömmlichen relationalen Datenbanken kann diese Leistung nicht erbracht werden.

Big Data

Dimensionen

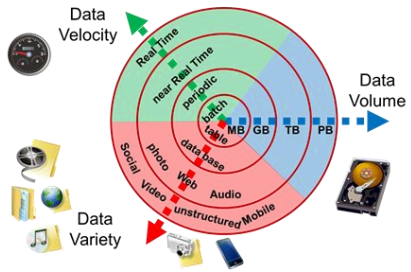


Data Veracity

- Betrifft die Unsicherheit (Wahrhaftigkeit) der Daten und die Datenqualität.
- Daten kommen aus verschiedensten Quellen, besitzen teilweise nicht die gewünschte Qualität.
- Für den eigentlichen Einsatzzweck müssen die Daten überprüft und aufbereitet werden.

Big Data

Dimensionen



Data Value

- Value = der Mehrwert oder Business Value
- Daten haben einen immanenten Wert, sie werden auch als Gold des Digitalzeitalters bezeichnet.
- Der große Marktwert der größten Technologieunternehmen der Welt kommt von den Daten, die sie ständig, für mehr Effizienz und neue Produkte, sammeln und analysieren.
- Dabei geht es nicht nur um Analyse, sondern darum den Wert der Daten zu erkennen, durch Mustererkennung, treffen fundierter Annahmen und Verhaltensvorhersagen.

Big Data

Anwendungszwecke

- Recruiting durch datengestützte Webanalyse
- Betrugsprävention und Einhaltung von Vorschriften
- Vorhersage von Epidemien
- Produktentwicklung
- Risikobewertung bei Versicherungen
- Erstellung von unterschiedlichen Profilen (Bewegung, Kaufverhalten usw.)
- Derzeit entstehen fast täglich neue Einsatzgebiete für Big Data und in deren Kontext genannte Teilgebiete (z.B. predective Analytics)

Big Data

Risiken

- Big Data als der datengestützte Weg in den totalitären Überwachungsstaat?
- Der gläserne Angestellte
- Abhängigkeit von Algorithmen
- Aushöhlung des Datenschutzes

ETL

ETL (Extract, Transform, Load)

- Mit ETL wird ein Prozessbezeichnet, bei dem Daten aus mehreren unterschiedlich strukturierten Datenquellen in einer Zieldatenbank oder auch Data Warehouse vereinigt werden.
- Die drei Phasen des Prozesses:
    1. Extract: Extraktion der Daten aus verschiedenen Quellen
    2. Transform: Transformation der Datenstruktur und Dateninhalte in das Format und Schema der Zieldatenbank
    3. Load: Laden der Daten in die Zieldatenbank

ETL (Extract, Transform, Load)

Extraktion

- Bei diesem Schritt werden Teilbereiche der Daten aus den verschiedenen Quellen extrahiert und für die Transformation vorbereitet.
- Die Extraktion kann synchron oder asynchron erfolgen.
- Asynchrone Extraktion kann periodisch, ereignis- oder anfragegesteuert erfolgen.

ETL (Extract, Transform, Load)

Transfomation

- Bei der Transfomation werden die extrahierten Daten an das Format und das Schema der Zieldatenbank angepasst.
- Der Transformationsprozess besteht aus mehreren Einzelschritten:
  - Festlegung grundlegender Aspekte der Formatierung
  - Bereinigung fehlerhafter Daten
  - Prüfen auf ähnliche Informationen und Datenduplikate
  - Gruppieren, Sortieren und Aggregieren der Daten
  - Anpassung an Zielformate und Zielschemata

ETL (Extract, Transform, Load)

Laden

- Der letzte Schritt ist das Laden der zuvor geprüften und angereicherten Daten.
- Hier erfolgt die eigentliche Integration in die Zieldatenbank, die Daten werden physisch verschoben.
- Bei diesem Schritt ist die Integrität der Daten sicherzustellen.

Quellen

Buchquelle

Kersken, Sascha (2017): IT-Handbuch für Fachinformatiker. Der Ausbildungsbegleiter. 8. Auflage, revidierte Ausgabe. Bonn: Rheinwerk Verlag; Rheinwerk Computing.

Luber, Stefan (2018): Was ist ETL (Extract, Transform, Load)? In: BigData-Insider, 16.11.2018. Online verfügbar unter https://www.bigdata-insider.de/was-ist-etl-extract-transform-load-a-776549/, zuletzt geprüft am 28.06.2021.

Noack, Alexander (2020): | dm. In: CLICKHERO GmbH, 09.09.2020. Online verfügbar unter https://digital-magazin.de/definition-von-big-data/, zuletzt geprüft am 28.06.2021.

Wuttke, Laurenz (2020): Was ist Big Data? Definition, 4 V's und Technologie. In: datasolut GmbH, 29.04.2020. Online verfügbar unter https://datasolut.com/was-ist-big-data/, zuletzt geprüft am 28.06.2021.

Abbildungen

1 „Big_Data“ Lizenz: Ender005 (https://commons.wikimedia.org/wiki/File:Big_Data.png), https://creativecommons.org/licenses/by-sa/4.0/legalcode

VIELEN DANK!