Big Data

Jonny Rillich

24. Juli 2015

1 Motivation

Durch die Erfolgsgeschichte des Internets ist eine Vielzahl von Anwendungen entstanden, welche einen Anteil am täglichen Leben von Millionen von Nutzern spielt. Portale wie Facebook, Twitter und Youtube erfreuen sich stetig an steigenden Nutzerzahlen. "Pro Minute werden zum Beispiel über Google mehr als zwei Millionen Suchanfragen abgesetzt, über Amazon mehr als 80000 Dollar umgesetzt oder in YouTube 30 Stunden Videomaterial hochgeladen und 1,3 Millionen Videos konsumiert."[1] Durch stetige Protokollierung, Persistierung und Dublizierung wächst das weltweite Datenwachstum jährlich um etwa 30%, was einer Verdopplung aller 2,5 Jahren entspricht. Schätzungen zufolge soll im Jahr 2020 das weltweite Datenvolumen auf 5 Zetabyte angestiegen sein.

In diesen Datenmassen schlummert ein riesiges finanzielles sowie wissenschaftliches Potenzial. Doch um dieses Potenzial ausschöpfen zu können, werden Alternativen zu herkömmlichen Datenbankverarbeitungssystemen benötigt, da diese nicht für solche Mengen an Daten ausgelegt sind. Big Data beschreibt demnach die Daten, welche zu groß und unstrukturiert sind, um von herkömmlichen Systemen verarbeitet zu werden.

2 Grundlagen

"Der Ursprung und die erstmalige Verwendung des Begriffes Big Data im aktuellen Kontext sind nicht ganz eindeutig und es werden unterschiedliche Quellen genannt, die den Begriff in der aktuellen Verwendung geprägt haben könnten. Relativ unumstritten jedoch ist die Definition der Eigenschaften von Big Data durch Gartner im Jahr 2011"[1]. Dieser beschreibt Big Data durch das 3-V-Modell, welches die Schwierigkeiten des Datenwachstums in 3 Dimensionen darstellen lässt. Das ursprüngliche Modell wurde in einigen Veröffentlichungen um eine weitere Dimension, Value, erweitert[2].

2.1 V-Modell

2.1.1 Volume

- Beschreibt die Masse an Daten

2.1.2 Velocity

- Beschreibt die enorme Datenrate, mit der aus unterschiedlichsten Daten erzeugt werden
- Beschreibt außerdem die Geschwindigkeit, in welcher die erzeugten Daten verarbeitet werden

2.1.3 Variety

- Beschreibt die Vielzahl an unterschiedlichen, oft unstrukturierten Datentypen

[1] http://www.gi.de/service/informatiklexikon/detailansicht/article/big-data.html
[2] https://en.wikipedia.org/wiki/Big_data

2.1.4 Value

- Beschreibt den Wert der Daten
- darunter fallen Aktualität, Struktur, Vollständigkeit etc.

3 Anwendungsbereiche

Es gibt unzählige Anwendung aus unterschiedlichen Bereichen wie Wirtschaft, Wissenschaft oder Medizin, in denen es sich lohnt mit Big Data zu beschäftigen.

Als Pionier für die Verarbeitung von Big Data gilt der amerikanische Informatiker Jim Grey, welcher es sich mit anderen Forschern im Projekt "Sloan Digital Sky Survey" zur Aufgabe gemacht hat, den Himmel digital zu erfassen. Das Projekt begann im Jahr 2000 und Endete 2005. In dieser Zeit wurden 930.000 Galaxien und 120.000 Quasare digitalisiert, dabei wurden täglich 250 Gb Daten produziert, was zu damaligen Zeit enorm war. Heutige Anwendungen Verarbeiten diese Datenmengen problemlos in kürzester Zeit. Dadurch können z.B hochkomplexe Anfragen auf Millionen Datensätze ausgewertet werden.

Aus wirtschaftlicher Sicht, geht der Begriff Big Data oft einher mit der Aussage, "Big Data ist das Öl der Neuzeit". Dies bedeutet, dass die Daten in unverarbeiteter Form relativ nutzlos sind, wenn es jedoch gelingt, "durch aufwändige Verfahren und Analysen Struktur in die Daten zu bekommen, dann können sie zur Beantwortung von neuen Fragestellungen genutzt werden und ihr finanzielles Potential entfalten"[1].

4 Komponenten

4.1 Anforderungen

Das V-Modell beschreibt definiert die Eigenschaften von Big Data aus technischer Sicht. Um die Sicht der Anforderungen des Nutzers an eine Big Data Anwendung wiederzuspiegeln, wird das F-Modell eingeführt. Dieses beinhaltet die Attribute: Fast, Felxible, Focused.

4.1.1 F-Modell

Fast:

- Die Anwendung soll das benötigte Ergebnis schnell liefern

Flexible:

- Die Anwendung soll sich ohne großen Aufwand an veränderte Bedingungen anpassen lassen
- z.B. das Einbeziehen neuer Datenquellen oder Veränderung der Datenstruktur

Focused:

- Die Anwendung soll in der Lage sein, relevante Datenquellen selektieren zu können

4.1.2 Komponenten

5 Technologien

Um Big Data zu verarbeiten, reicht die Geschwindigkeit von herkömmlichen relationalen Datenbanken nicht aus. Neue Datenbanktechnologien, welche Daten im Terabyte- oder sogar Peta-Bereich verarbeiten können, werden No-SQL(Not only-SQL) Datenbanken bezeichnet. Diese werden in die Bereiche Key-Value-Stores, Dokumentenorientierte Datenbanksysteme, Spaltenorientierte Datenbanksysteme und Graphen Datenbanken kategorisiert.

5.1 Key-Value-Stores

Key-Value-Stores bedienen sich eines einfachen Schlüssel-Wert Schemas, analog zu einer Hashtabelle. Dieser Aufbau bringt den Vorteil einer extrem hohen Skalierbarkeit und hat einen effiziente Datenverwaltung zur Folge. Nachteilig ist, dass keine komplexen Abfragen an das Datenbanksystem gestellt werden können.

5.2 Dokumentenorientierte Datenbanksysteme

In einer Dokumentenorientierten Datenbank, werden die Daten in Dokumenten abgespeichert. Anders als bei herkömmlichen Datenbanksysteme unterliegen diese Daten keinem Schema. Es ist möglich, dass jeder Dateneintrag in der Datenbank eine andere Struktur besitzt. Es ist es möglich, vorhandenen Dokumenten weitere Dokumente hinzuzufügen, dadurch können komplexe verschachtelte Strukturen aufgebaut werden.

5.3 Spaltenorientierte Datenbanksysteme

Ein spaltenorientiertes Datenbankssystem speichert Datenbankeinträge in Spalten statt in Zeilen. Diese Art der Speicherung hat den Vorteil, dass nur die benötigten Attribute einer Relation ausgelesen werden müssen, anstatt alle auszulesen. Müssen stattdessen alle Attribute einer Relation ausgelesen werden, ist diese Art der Speicherung langsamer, da die Attribute physisch verteilt auf der Datenbank liegen können und so unter Umständen nicht sequentiell ausgelesen werden können.

5.4 Graphen Datenbanken

Graphen Datenbanken speichern Knoten und Kanten. Jeder Knoten kann dabei mehrere Kanten besitzen. Kanten können Beziehungen zwischen Knoten abbilden. Dadurch das jeder Knoten beliebig viele Kanten besitzen kann, ist es möglich multidimensionale Beziehungen zwischen den Knoten darzustellen. Durch seine Struktur ist es möglich schnell durch den Graphen zu traversieren, was unter relationalen Datenbanken nur durch rechenaufwändige Join-Operationen zu realisieren wäre.