
Big Data Grundlagen

Inhaltsverzeichnis

Big Data Hype

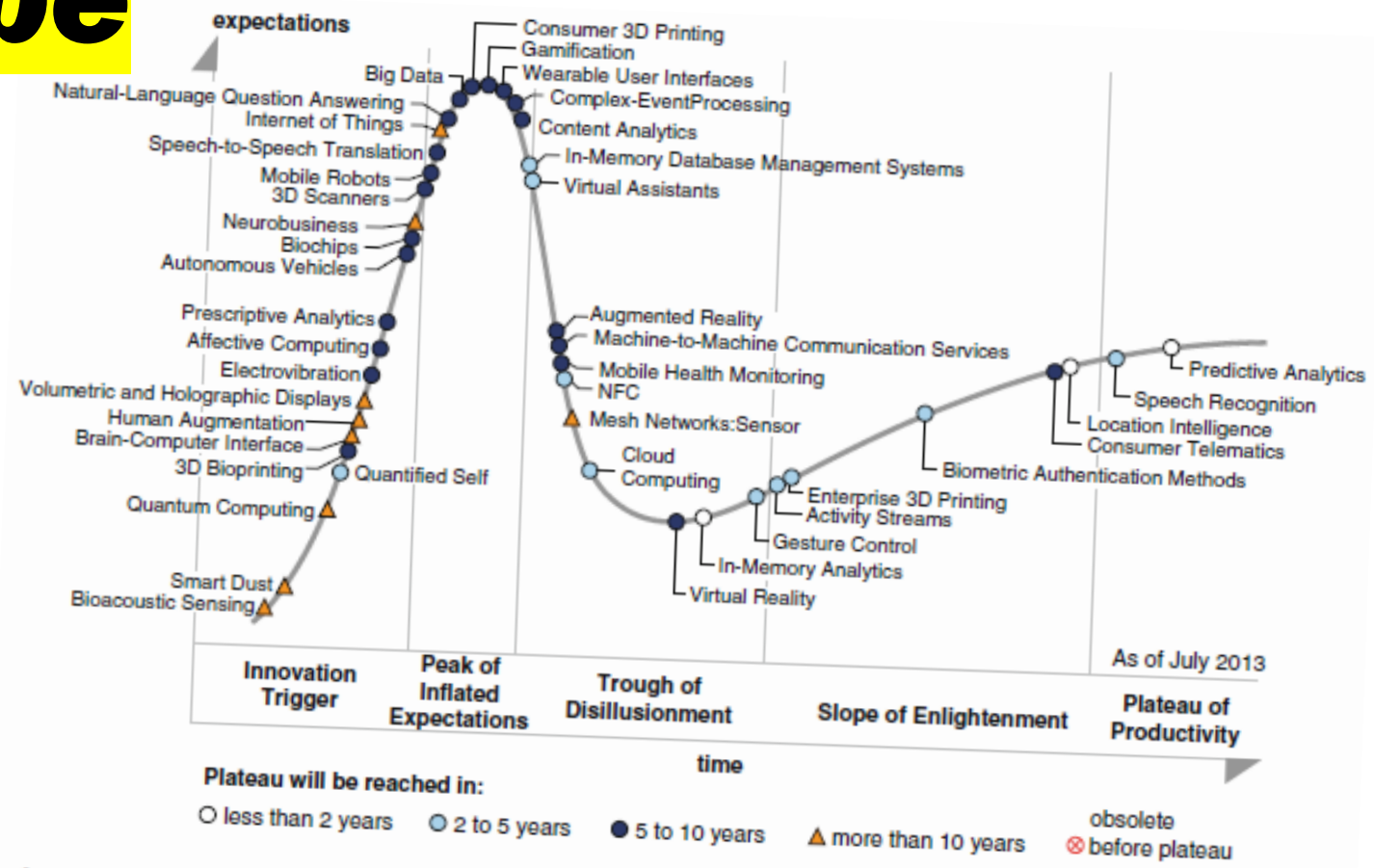

Abb. 1.1 Gartners Hype-Zyklus von 2013 (http://www.gartner.com/newsroom/id/2575515)

Big Data

- Daten, die in ihrer Größe die klassische Datenhaltung, Verarbeitung und Analyse auf konventioneller Hardware übersteigen
- Heterogener als klassische Daten
- Unternehmensdaten werden um externe Daten erweitert

  \+ Erweiterte Sicht auf das Unternehmen
- Erhöhung der Informationstransparenz und Frequenz zur Verarbeitung und Analyse von Daten

Die 5 V's

- Volume: Umfangreicher Datenbestand (Tera- bis Zettabytebereich)
- Variety: Speicherung von strukturierten, semi-strukturierten und unstrukturierten Multimedia-Daten
- Velocity: Geschwindigkeit – Auswertung und Analyse der Datenströme in Echtzeit

Ergänzt durch:

- Value: Steigerung des Unternehmenswertes
- Veracity: Aufrichtigkeit - Berücksichtigung der unterschiedlichen Datenqualität der Datenbestände

Relationenmodell

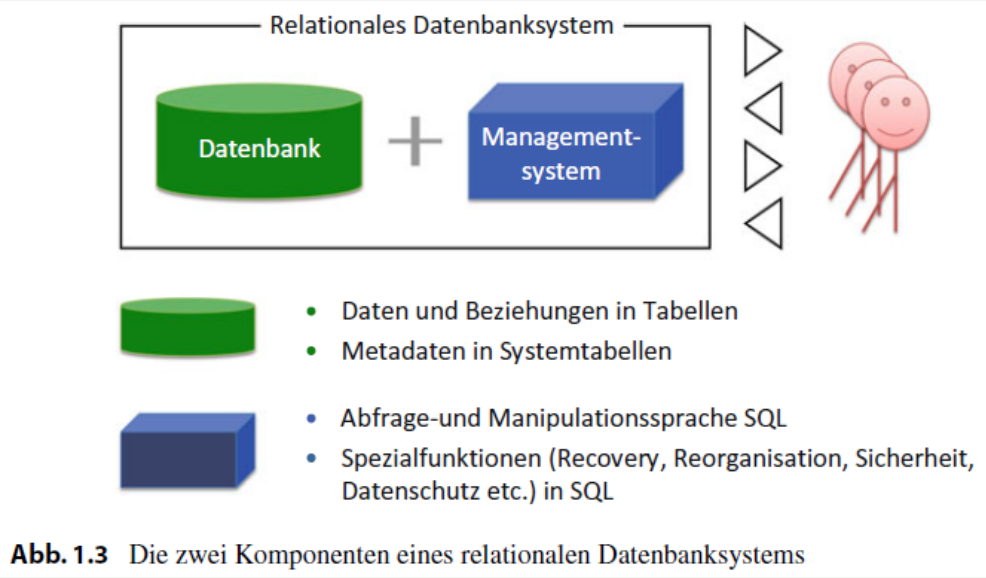

Abb. 1.3 Die zwei Komponenten eines relationalen Datenbanksystems

- Speicherungskomponenten: Ablage von Daten und Beziehungen in Tabellen
  - \+ Tabellen mit den eigentlichen Daten
  - \+ Systemtabellen
- Verwaltungskomponenten: Datendefinitions-, Datenselektions- und Datenmanipulationssprache SQL sowie Dienstfunktionen für die Wiederherstellung

Nicht-relationale Datenbanken

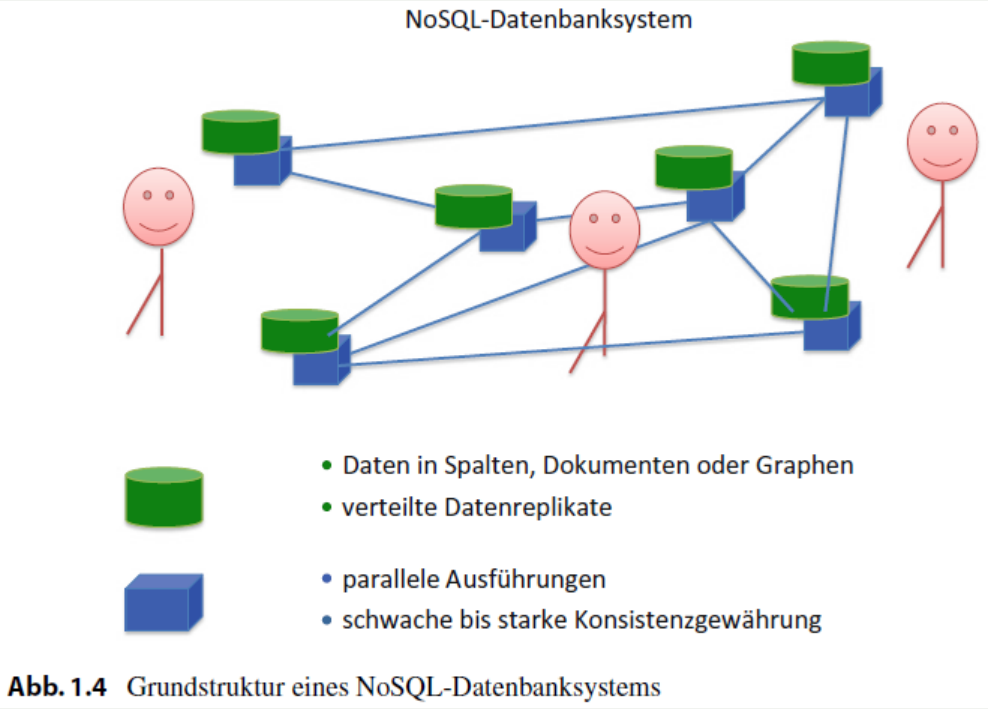

Abb. 1.4 Grundstruktur eines NoSQL-Datenbanksystems

- Massiv verteilte und replizierte Datenhaltungsarchitektur
  - + Datenreplikation wird unterstützt
  - + Konsistenzregeln sind konfigurierbar
- Vier Datenmodelle
  - + Key/Value Store
  - + Column Store
  - + Document Store
  - + Graph Database

Nicht-relationale vs. relationale Datenbanken

Relationale Datenbanken

- Immer konsistente Daten
- Vertikale Skalierung
- Verteilung auf mehrere Maschinen erhöht Komplexität
- Hohe Hardwareanforderungen und Hardwarekosten
- Datenvolumen größer 100 TB schwer zu verarbeiten
- Vielfalt von Strukturen erhöht Komplexität

Nicht-relationale Datenbanken

- Häufig bei webbasierten Firmen
- Verteilt und hochverfügbar aber nicht immer konsistent
- Sehr einfache horizontale Skalierung
- Verteilung auf mehrere Maschinen problemlos
- Günstige Hardware einsetzbar
- Dadurch große Datenvolumen möglich
- Verteilung ermöglicht Parallelisierung (Map/Reduce-Verfahren)

-> Relationale Datenbanken sind nicht geeignet für Big Data

Nicht-relationale vs. Relationale Datenbanken

	NoSQL	SQL
Modell	Nicht relational	Relational (Tabellen)
Architektur	Unterstützung verteilter Webanwendungen und horizontaler Skalierung	Datenunabhängigkeit - Daten und Anwendungsprogramme bleiben getrennt voneinander
Schema	Kein festes Datenbankschema	Relationales Datenbankschema
Mehrbenutzerbetrieb	Mehrbenutzerbetrieb ist möglich	Mehrbenutzerbetrieb ist möglich
Konsistenzgewährung	CAP-Theorem – Konsistenz nachrangig nach hoher Verfügbarkeit und Ausfalltoleranz	Bereitstellung von Hilfsmittel zur Gewährleistung der Datenintegrität

Berufsfelder

- Datenarchitekt
  - Verantwortlich für Datenarchitektur
  - Entscheiden in welcher Form Datenbestände bereitgestellt werden
- Datenbankspezialist
  - Beherrschen Datenbank- und Systemtechnik
  - Verantwortlich für physische Auslegung der Datenarchitektur
  - Entscheidung über den Einsatz der Datenbanksysteme
  - Zuständig für Verteilungskonzept, Archivierung und Restaurierung der Datenbestände
- Data Scientist
  - Spezialisten des Business Analytics
  - Datenanalyse und –interpretation
  - Wissensgenerierung
  - Data Mining, Statistik und Visualisierung

Datenmanagement mit SQL und NoSQL

Problematik

- Strukturierte Daten : Relationales Datenbankmodel
- Unstrukturierte Daten : ???

- Mit Entwicklung des Webs haben sich NoSQL Ansätze bewährt:
  - Strukturierte
  - Semi-strukturierte
  - Unstrukturierte
  - Echtzeitverarbeitung

Beispiel Webshop

- SQL:
  + Relationale Datenbank
- NoSQL:
  + Key/Value Store
  + Document Store
  + InMemory Datenbank
  + Graph Datenbank

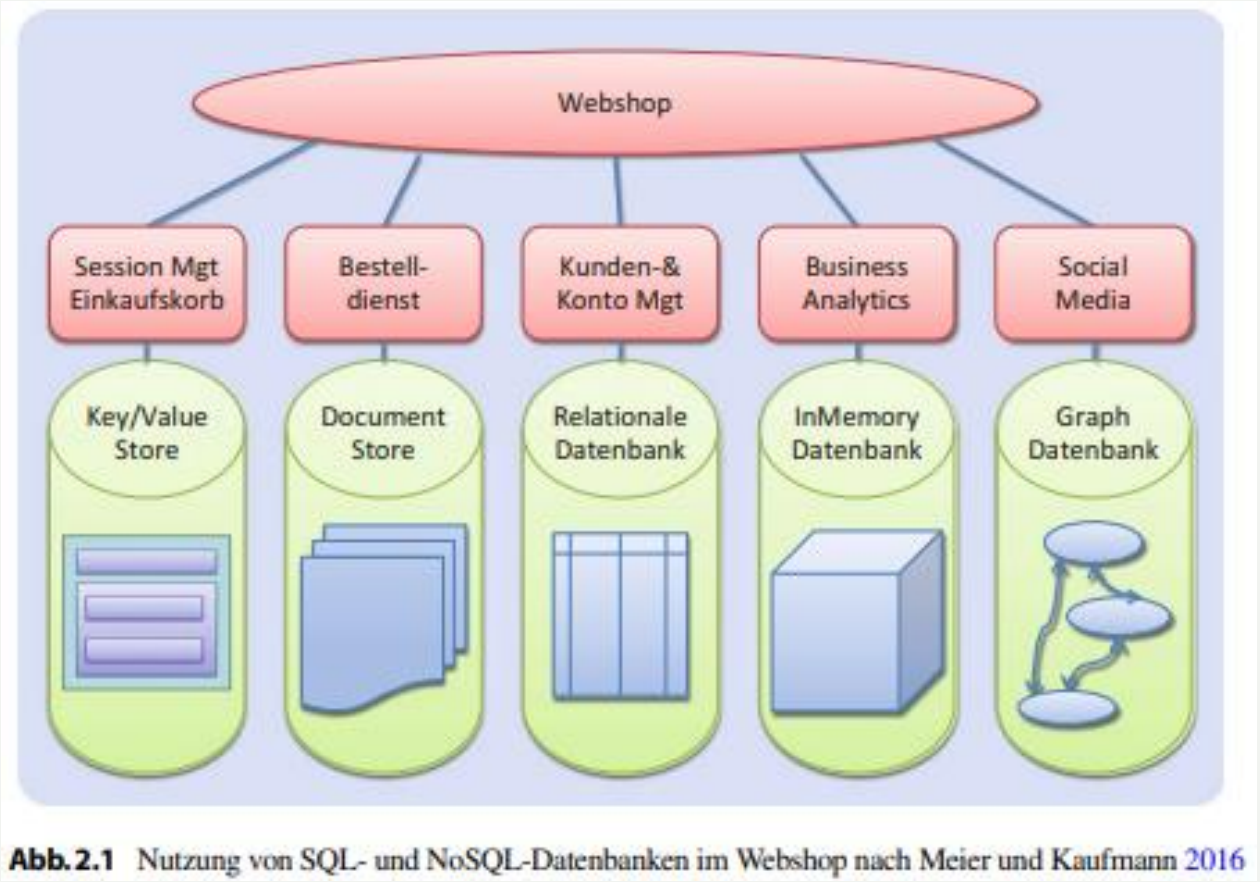

Abb. 2.1 Nutzung von SQL- und NoSQL-Datenbanken im Webshop nach Meier und Kaufmann 2016

Semantische Modelbildung

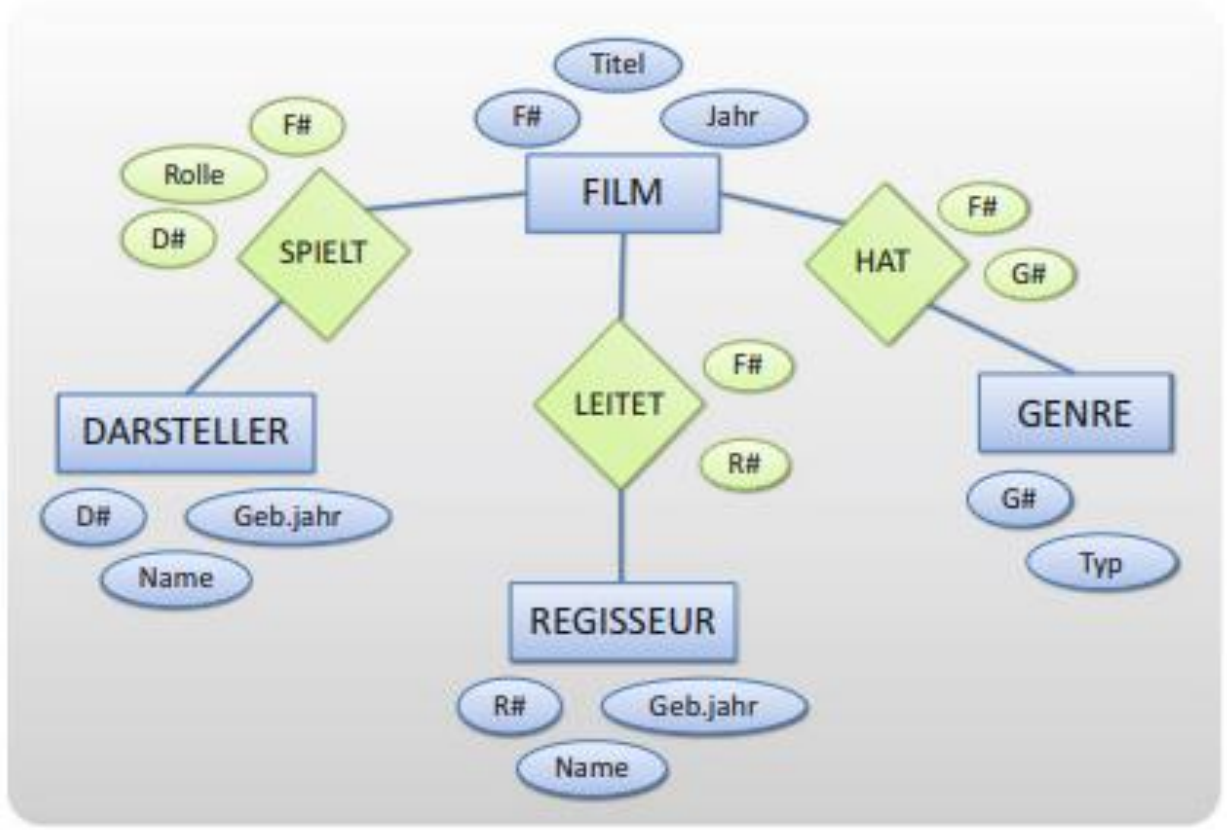

Abb. 2.2 Entitäten-Beziehungsmodell für Filme und Filmemacher

- Abstraktion der realen Welt in ein semantisches Datenmodel

Relationen-modell

- Abfragesprache SQL (Structured Query Language)
- Tabellen (Entitäten und Relationen)
  mit Spalten (Attribute)
  und Tupeln (Datensätze)

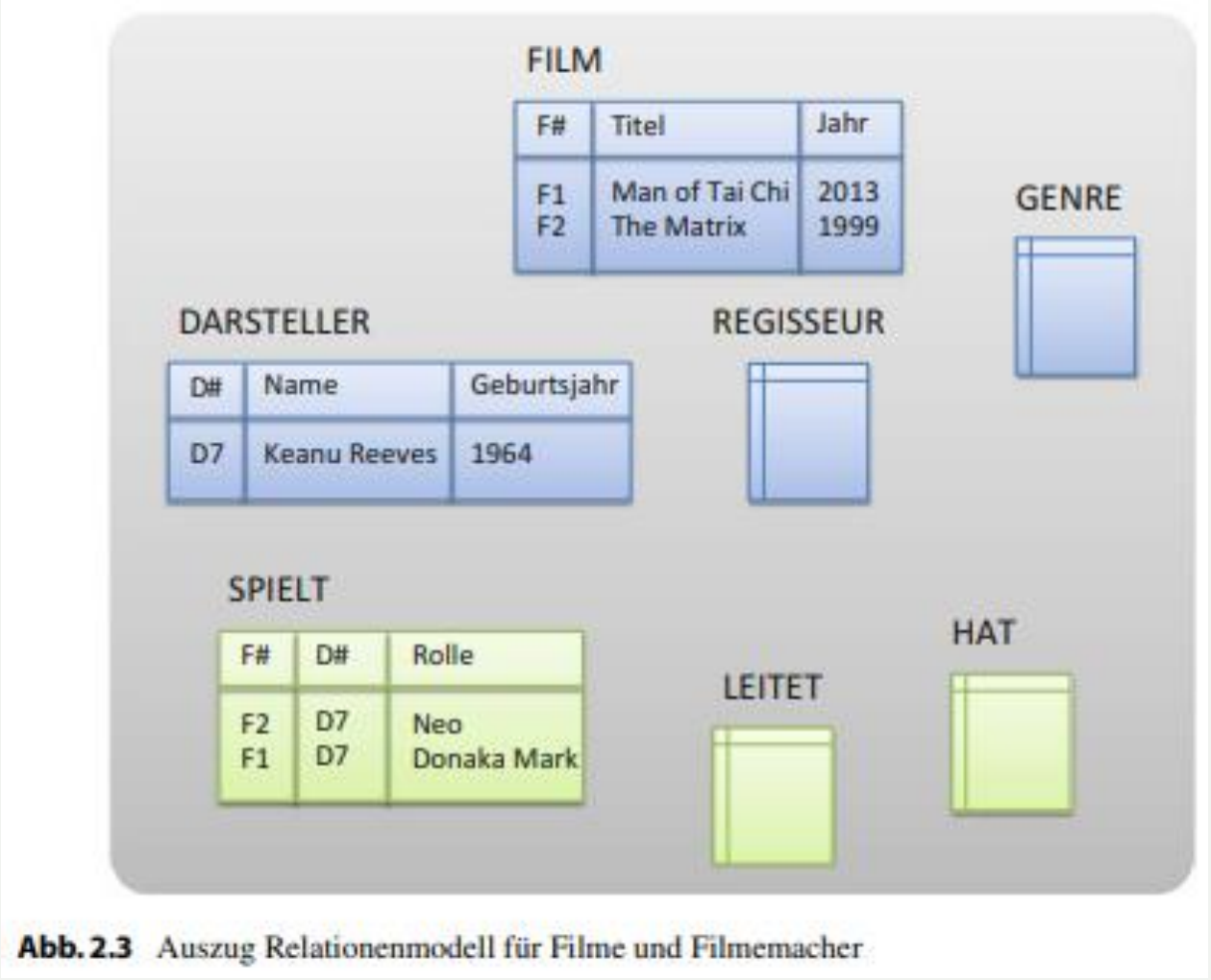

Abb. 2.3 Auszug Relationenmodell für Filme und Filmemacher

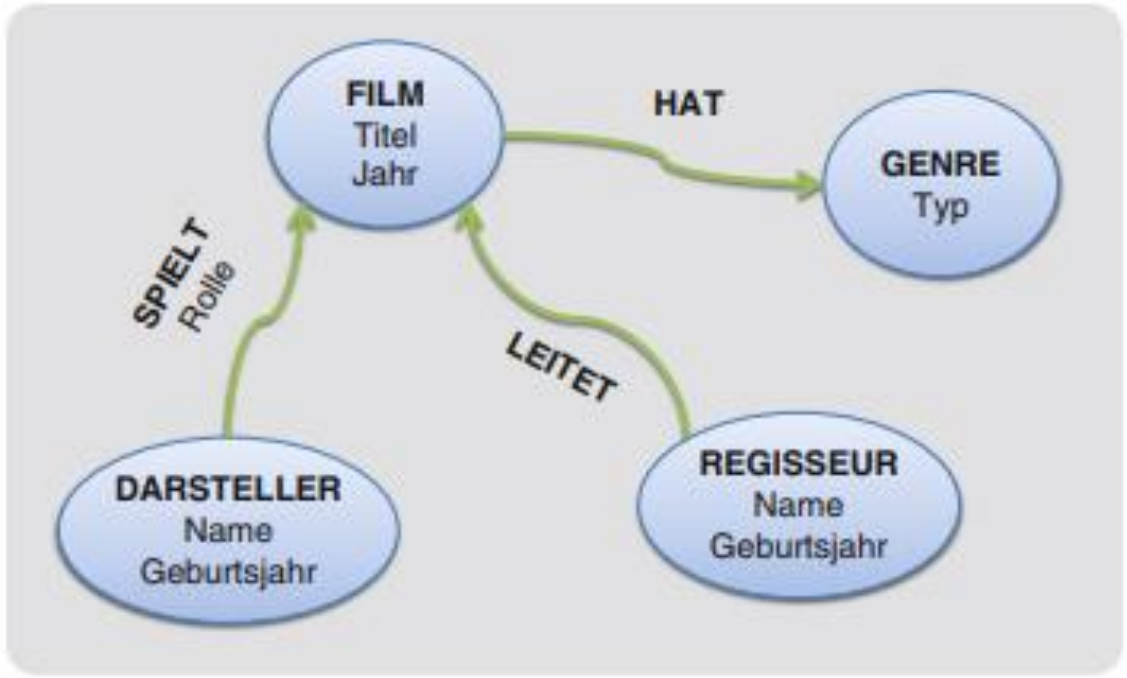

Abb. 2.4 Ausschnitt eines Graphenmodells für Filme und Filmemacher

Graphenmodel

- Analyse oder Optimierung netzwerkartiger Strukturen
- Dijkstra-Algorithmus (kürzeste Wegberechnung)

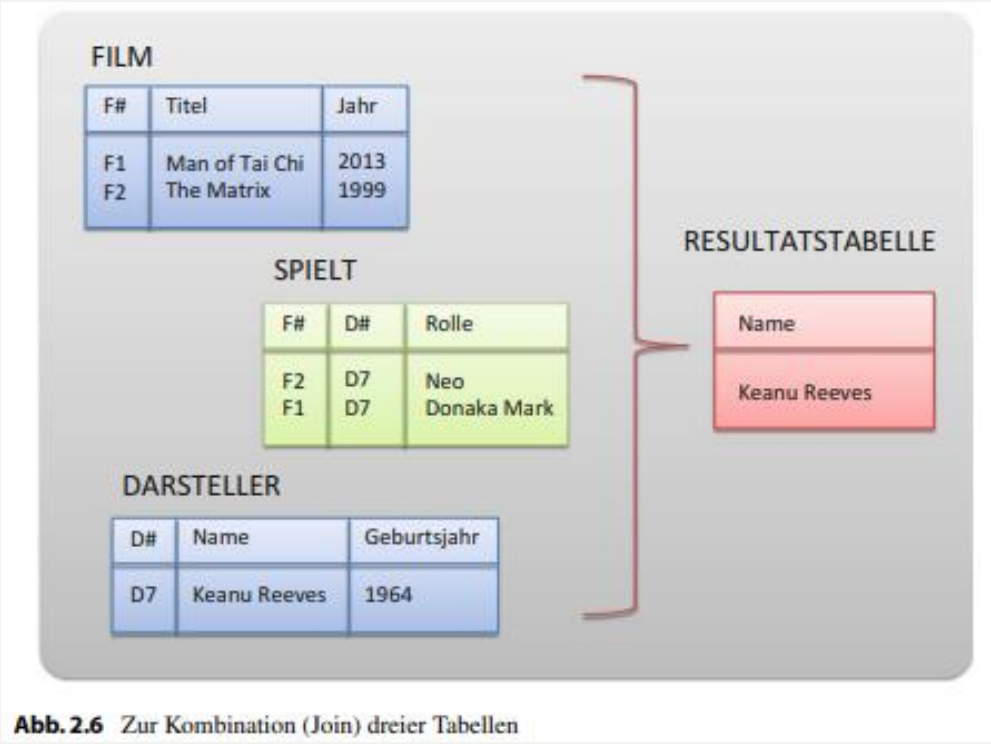

Abb. 2.6 Zur Kombination (Join) dreier Tabellen

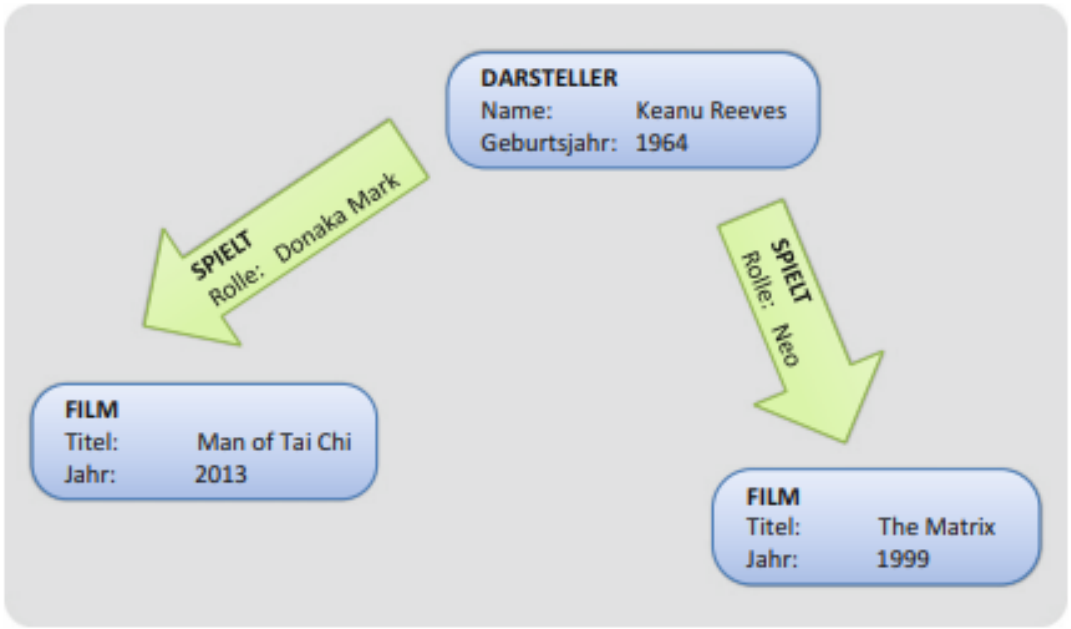

Abb. 2.7 Ausschnitt der Graphdatenbank über Keanu Reeves

Relationale Datenbank: SQL

```
SELECT Jahr
FROM FILM
WHERE Titel = ‚The Matrix‘;
```

Graphdatenbank (NoSQL): Cypher

```
MATCH (m: FILM {Titel: „The Matrix“})
RETURN m.Jahr
```

Abfragesprachen

Konsistenzgewährung

- Konsistenz beschreibt den Zustand widerspruchsfreier Daten
- Problematik: Mehrere Benutzer greifen gleichzeitig auf die Datenbank zu
- Konsistenzforderung bei umfangreichen Systemen nicht immer zu fordern

Cap Theorem

- Massiv verteilte Systeme können nur 2 der 3 Eigenschaften erfüllen:
  + Konsistenz (C)
  + Verfügbarkeit (A)
  + Ausfalltoleranz (P)

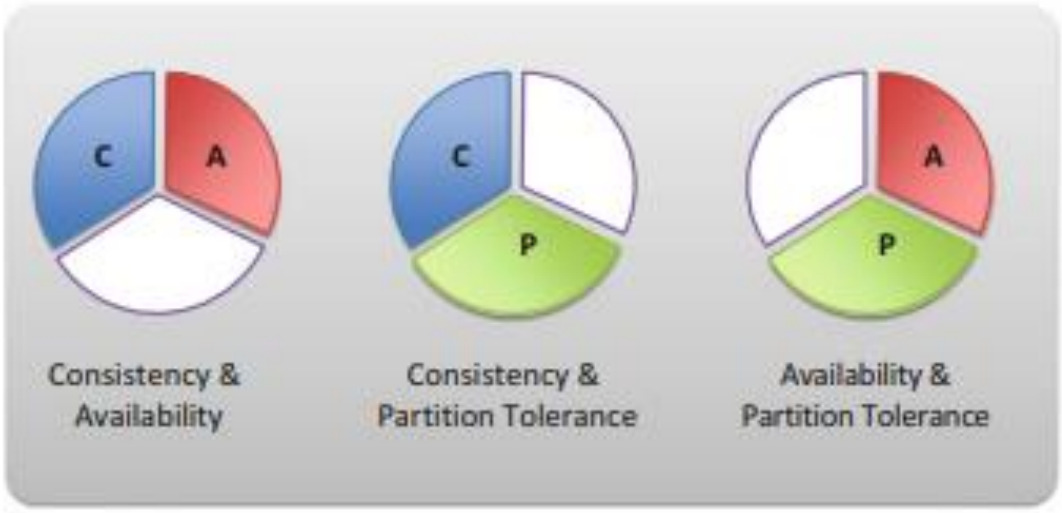

Abb. 2.10 Die möglichen drei Optionen des CAP-Theorems

Status Quo, Chance und Herausforderungen im Umfeld BI & Big Data

- Informationen als strategische Ressource
- Informationen sind wettbewerbskritisch

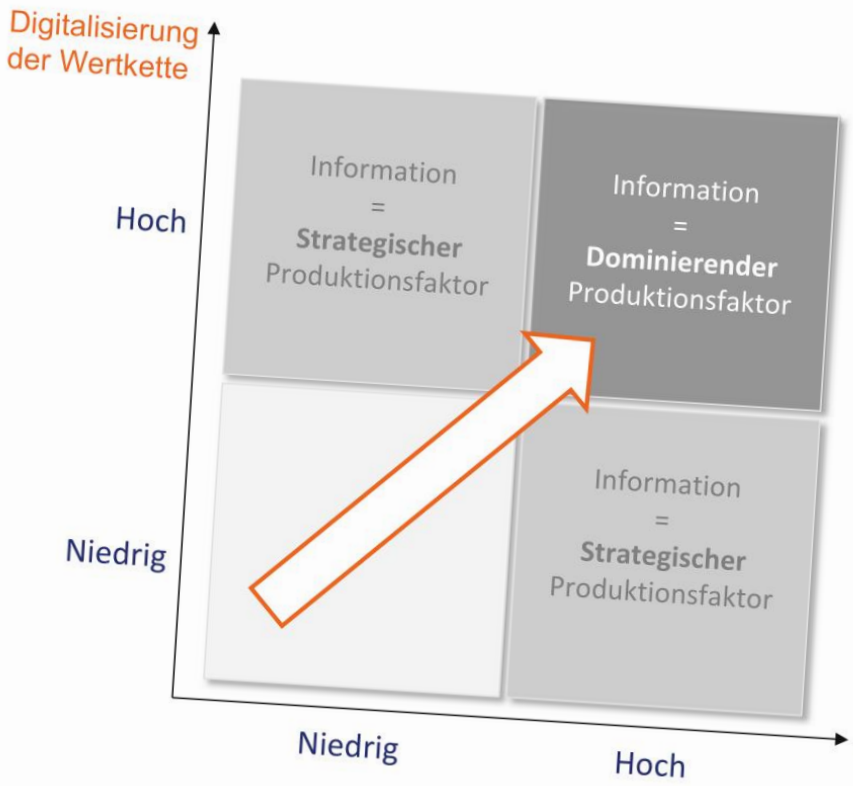

Abb. 3.1 Implikationen zunehmender Digitalisierung für Unternehmen

Daten

- Neue Datenmengen erkennen
- z.B.: User generated Content
- Daten aus Echtzeitvernetzung

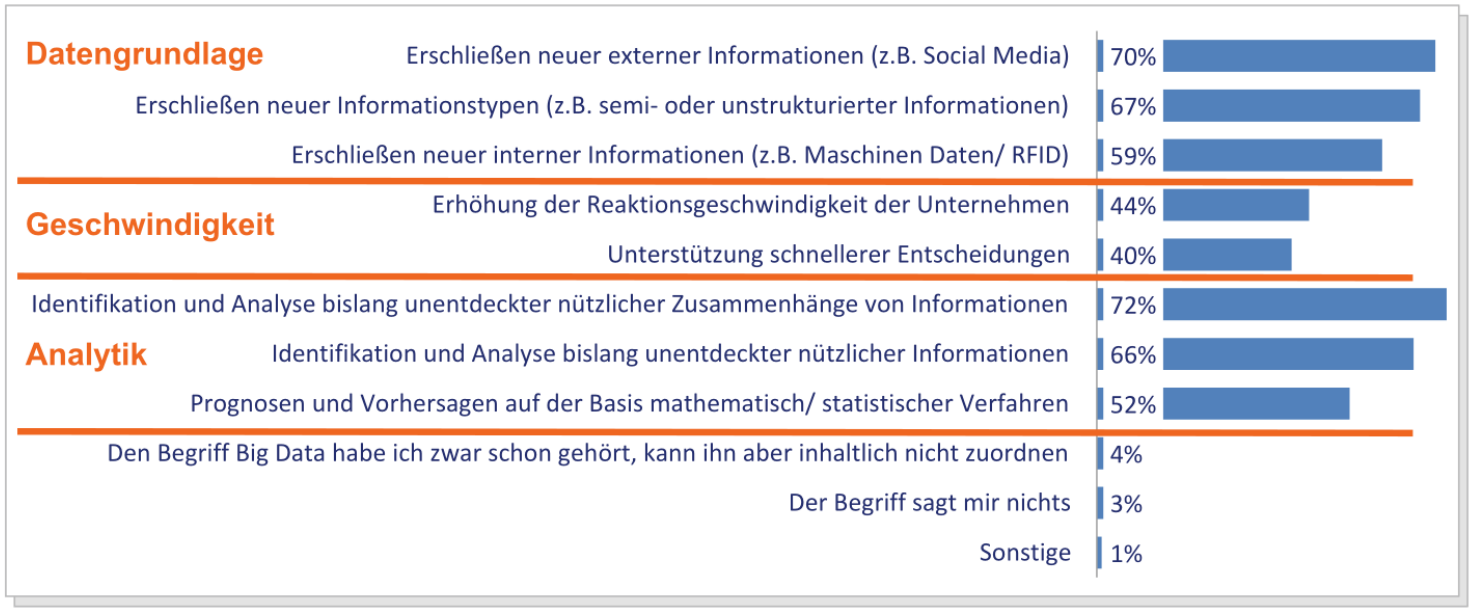

Abb. 3.2 Big Data Verständnis – Einzelkategorien

Studie zum Verständnis von Big Data

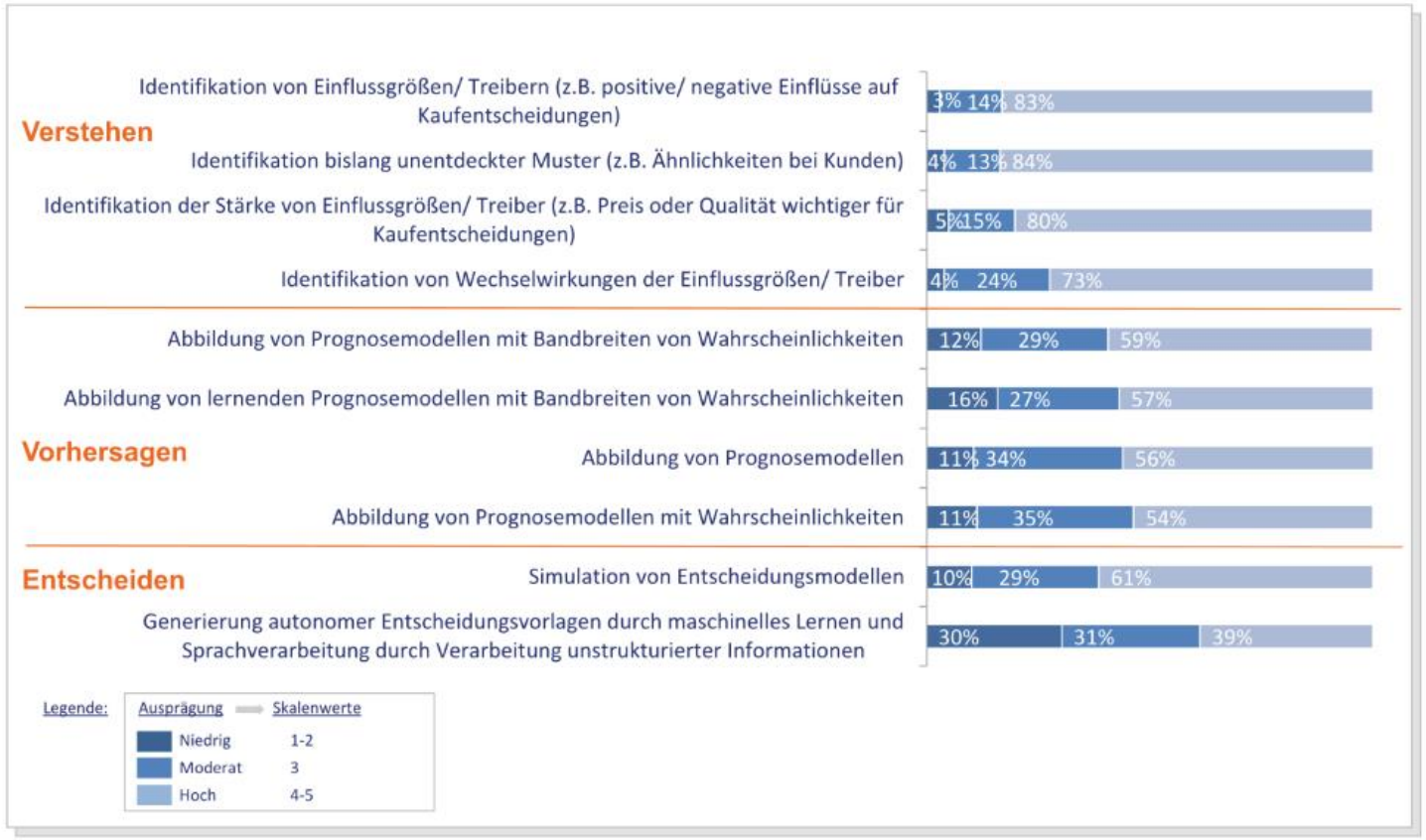

Abb. 3.4 Big Data – Analytische Potenziale

Häufige Verwendung zur Analyse / Verständnis sowie für Vorhersagen

Nutzungsbereiche

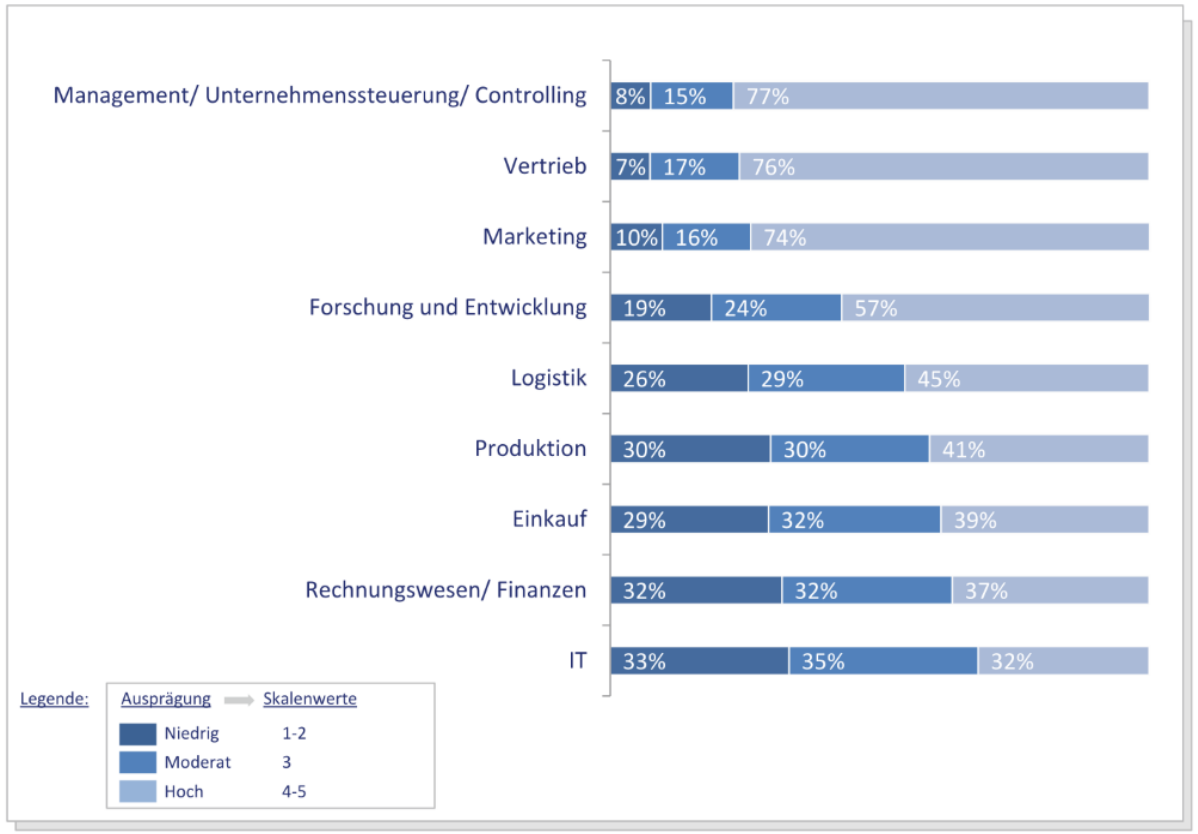

Abb. 3.5 Big Data – Analytische Potenziale in Funktionen

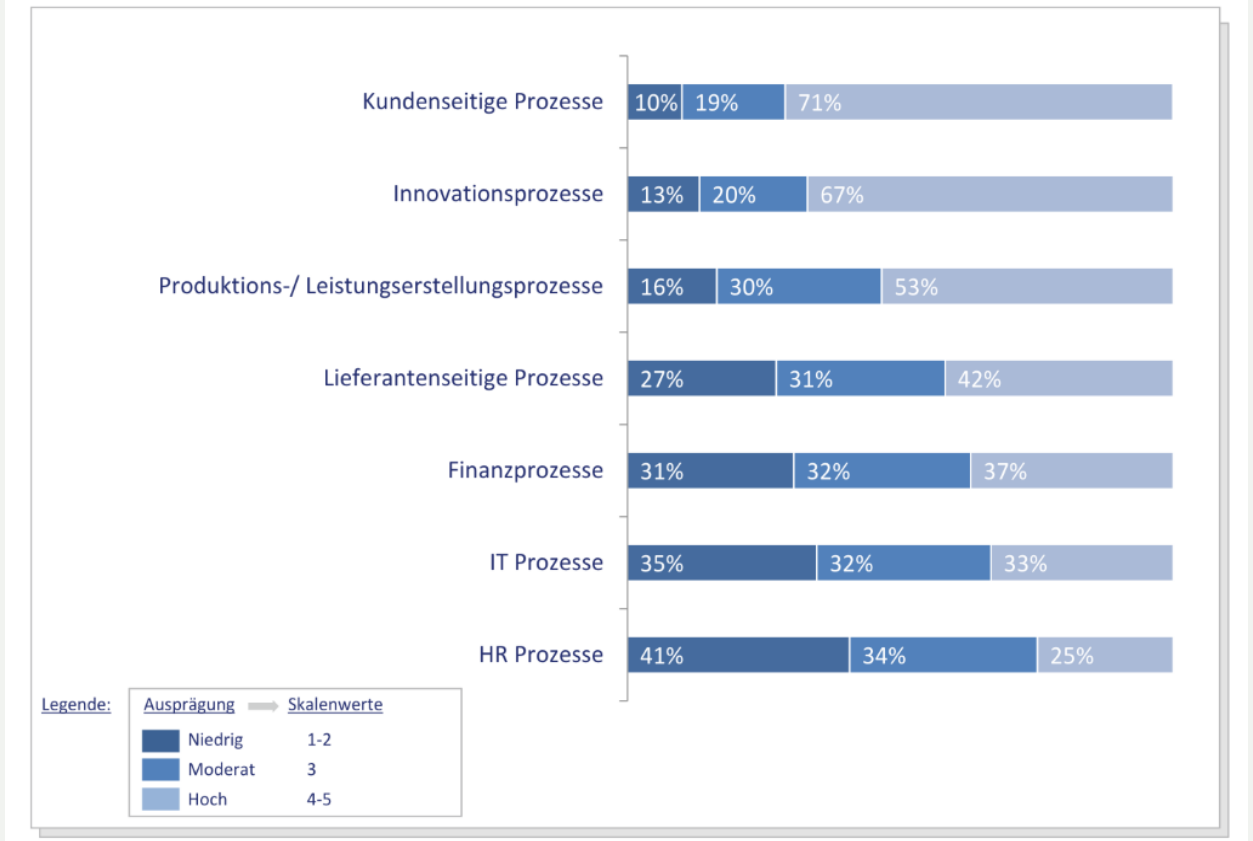

Abb. 3.6 Big Data – Analytische Potenziale in Prozessen

Nutzungspotential und Probleme

- z.B.: Industrie 4.0, Home Automation, Energieerzeugung, Automobilindustrie
- Hohes Potenzial durch die Erschließung neuer Datenquellen
- Zu wenig Nutzung von externen Quellen
- Ernsthaftes Problem mit der immer noch manuellen Datensammlung
- Zentrale Herausforderung ist fehlendes Know-How

Danke für die Aufmerksamkeit