Big Data Datenvisualisierung zur Optimierung der Unternehmensprozesse

Harald Konnerth, Gerold Wagner, Werner Wetzlinger

FH OÖ Studienbetriebs GmbH, Harald Konnerth, Wehrgrabengasse 1-3, A-4400 Steyr, AUSTRIA

KURZFASSUNG/ABSTRACT:

Das Thema Big Data steigert seit Jahren die Präsenz in den Medien und wird mit der fortschreitenden Vernetzung und den Zukunftsthemen Internet of Things und Industrie 4.0 weiter an Popularität gewinnen.

Big Data stellt Methoden und Technologien zur Verfügung um trotz exponentiellen Wachstums der Datenmengen deren Verfügbarkeit zu sichern und die Nutzung der darin enthaltenen Informationen zu ermöglichen.

Die Herausforderung von Big Data besteht darin, Unternehmen in die Lage zu versetzen, die vorhandenen Daten der unterschiedlichen Quellen zu identifizieren und zu verknüpfen. Die entstehende Datenbasis kann zur Identifikation von Effizienzsteigerungspotenzialen der Geschäftsprozesse und als Grundlage zur Identifikation von Innovationspotenzialen dienen. Die Integration in die Wertschöpfungskette kann auf vielfältige Arten erfolgen, und kann darüber hinaus den Prozessablauf ändern bzw. umkehren. Voraussetzung für die Ausschöpfung dieses Potenzials ist eine Analyse und Datenaufbereitung der Datenbasis, die durch Identifikation der passenden Visualisierung den Erkenntnisprozess entsprechend unterstützen kann.

1 EINLEITUNG

Big Data stellt Methoden und Technologien zur Verfügung um trotz exponentiellen Wachstums der Datenmengen deren Verfügbarkeit zu sichern und die Nutzung der darin enthaltenen Informationen zu ermöglichen.

Big Data kann mit drei Dimensionen beschrieben werden, die auf eine Veröffentlichung von Doug Laney zurückzuführen sind. Er bezeichnet die Herausforderungen, die auf Grund des Datenwachstums entstehen, als 3-V Modell (volume, velocity, variety) [1]. Die drei Dimensionen ergeben sich aus einem ansteigenden Volumen, aus der steigenden Geschwindigkeit, mit der Daten generiert und verarbeitet werden müssen und auf immer vielfältigere Quellen, aus denen strukturierte wie unstrukturierte Daten stammen. Das Modell wurde um die vierte Dimension Richtigkeit (veracity) [2] erweitert. Neue Technologien insbesondere im Bereich der NoSQL Datenbanken ermöglichen die Auswertung dieser steigenden Datenmassen [3].

Die Herausforderung von Big Data besteht darin, Unternehmen in die Lage zu versetzen, die vorhandenen Daten der unterschiedlichen Quellen zu identifizieren und zu verknüpfen. Die entstehende Datenbasis kann zur Identifikation von Effizienzsteigerungspotenzialen der Geschäftsprozesse und als Grundlage zur Identifikation von Innovationspotenzialen dienen. Voraussetzung für die Ausschöpfung dieser Potenziale ist eine Unterstützung des Erkenntnisprozesses durch eine hinreichende Visualisierung.

Datenvisualisierung ist zur Analyse eine neue und zunehmend genutzte Technologie im Big Data Umfeld, die ad-hoc und interaktiv Fragestellungen durch menschliche Interaktion analysiert. Der dynamische Analyseansatz besteht im Erkennen von Mustern durch das menschliche Auge und wird durch analytische Algorithmen unterstützt.

2 FORSCHUNGSFRAGE

Welche Visualisierungsmethoden sind für welche Unternehmensdaten aus welchen Quellen geeignet und wie kann dies mittels Big Data Technologien umgesetzt werden?

3 METHODIK

Die Methodik umfasste eine Erhebung von Daten aus Geschäftsprozessen, eine Analyse von Darstellungsmethoden, eine Marktanalyse von NoSQL Datenbanken, und die Umsetzung anhand eines Prototyps.

- **Datenerhebung aus Geschäftsprozessen**: Es können keine allgemeinen Massendaten definiert werden, da die Entstehung vom speziellen Unternehmenskontext abhängig ist. Es konnte jedoch mittels Literaturanalyse im Kontext eines Webauftritts allgemein vorhandenen Datenquellen ermittelt werden.
- **Analyse von Darstellungsmethoden**: Es erfolgte eine Literaturrecherche von möglichen Klassifizierungs-Kriterien der Methoden, auf deren Basis ein morphologischen Kastens erstellt wurde. Daran anschließend wurden Visualisierungsmethoden erhoben und klassifiziert.
- **Marktanalyse von NoSQL Datenbanken**: Die Verknüpfung unterschiedlicher Datenbasen (z.B. strukturierte und unstrukturierte Daten) ist mit NoSQL Datenbanken möglich, die recherchiert und deren Potenzial für die Speicherung als Basis für die Visualisierung erhoben wurde.
- **Prototyping**: Mittels einem Prototypen wurden die recherchierten Darstellungsmethoden auf die Tauglichkeit hin überprüft.

4 RESULTATE

4.1 Datenerhebung aus Geschäftsprozessen

In einer Studie zur Bedeutung von Big Data in Unternehmen und Organisationen gaben 49% der Befragten zur Frage, welche sie von den drei wichtigsten Zielsetzungen als die oberste Priorität einstufen, kundenorientierte Ziele als oberste Priorität an und 18% gaben als zweitgereihte Priorität die Optimierung betrieblicher Abläufe an [4]. Das Ziel der Unternehmen besteht im Verständnis der Wünsche und Verhaltensweisen von Kunden um das Kauferlebnis verbessern zu können [4]. In dieser Studie berichteten die Teilnehmer, dass die primäre Datenquelle Systeme im eigenen Unternehmen sind. Nahezu drei von vier Befragten in Unternehmen mit laufenden Big Data-Projekten analysieren Log-Daten.

Der Zugang zu externen Daten stellt für Unternehmen eine Herausforderung dar. Daten können zwar von spezialisierten Anbietern zugekauft werden, jedoch bieten diese nicht alle gewünschten Datensammlungen an. Darüber hinaus besteht die Unsicherheit bzgl. der Rechte an externen Daten. Aus diesen Gründen wird die Erstellung einer unternehmensinternen Datenbasis als bedeutend angesehen [5]. Eine Herausforderung besteht auch darin nur die relevanten Daten, aus denen Wissen generiert werden kann, zu speichern um dem EU-Prinzip der Datenminimierung gerecht zu werden [6].

Unternehmen sehen vermehrt Social Media Daten als wichtige Ergänzung des bestehenden Datenpools an, wobei es unterschiedliche Auffassungen der Aussagekraft gibt, da nicht alle Kundengruppen Online Netzwerke nutzen [7]. Unsicherheit in der Datengenauigkeit besteht auch auf Grund von der Möglichkeit von unvollständigen, mehrdeutigen und inkonsistenten Daten [8].

Traditionelle Organisationen arbeiten nach der Wertschöpfungskette nach Porter, indem die Entwicklung und Innovation auf Basis von Marktforschung beruht, und anschließend die Produktion und Vermarktung durchgeführt wird. Mit dem Einsatz von Big Data Analysen wird der Prozess zum Teil umgekehrt (Abbildung 1) und die Anforderungen des Marktes werden bereits vor dem Innovationsprozess berücksichtigt. Daraus entstehen kundenorientierte und bei Bedarf individualisierte Produkte, die am Markt größere Erfolgswahrscheinlichkeit haben [9].

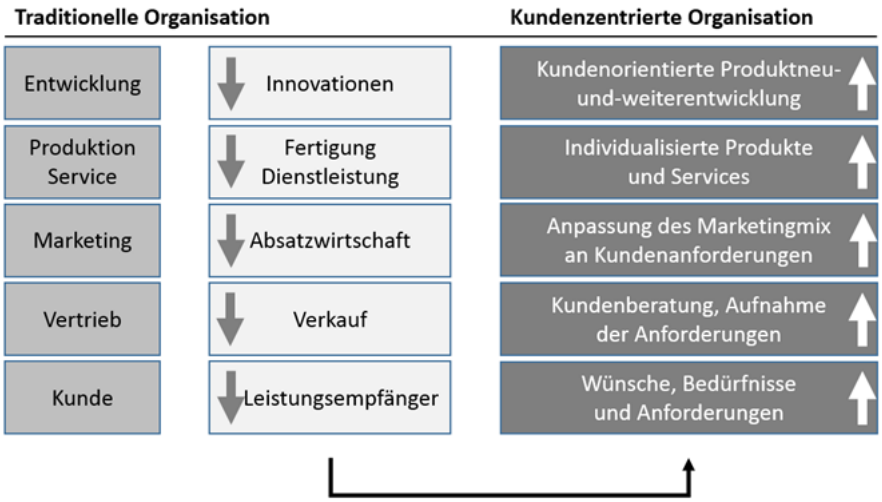

Abbildung 1. Umkehrung der Wertschöpfungskette (eigene Darstellung in Anlehnung an [9])

Eine Umkehrung der Wertschöpfungskette setzt das Wissen über die Wünsche, Bedürfnisse und Anforderungen des Marktes voraus, das aus Social Media Beiträgen gewonnen werden kann. Um die Marktanforderungen zu präzisieren bietet sich die Erweiterung der Analyse nach Standort an und nach demografischen Ausprägungen gliedern. Eine interaktive grafische Datenrepräsentation unterstützt die Interpretation der Informationen durch verschiedene Sichtweisen auf die Daten und hilft beim Erkennen von Mustern. Dies bildet auch die Voraussetzung der zeitnahen Umsetzung von Wünschen für die jeweilige Zielgruppe bis hin zur Identifizierung von Markttrends.

4.2 Analyse von Darstellungsmethoden

Die Recherche zu den Daten aus dem Webshop-Auftritt ergab folgende Datenquellen[4]: Webserver Log Datei, Webshop Log, Web Click Stream, Daten aus Blogs, Foren und Tweets und Bewertungsportale.

Die Vielzahl der **analysierten Darstellungsmethoden** unterscheidet sich im Wesentlichen in ihrer Dimensionalität, der Darstellungsform, der grafischen Aufbereitung und der Zeitpräsentation. Der entwickelte morphologische Kasten ermöglicht es die Methoden zu bewerten und zu vergleichen.

Tabelle 1. Darstellungsmethoden – morphologischer Kasten

Kriterium	Ausprägung				
Darstellungsdimension	1D	2D	3D		
Darstellungsform	Schrift	Tabelle	Diagramm	Zeichnung	Bild
Skalierung	leine	linear	logarithmisch	exponentiell	kategorisierend
Zeitrepräsentation	keine	diskret	kontinuierlich		
Interaktion	keine	Navigation i der Präsentation	mit dem grafischen Modell	mit dem Datenmodell	
Sicht	Einzelansicht	Mehrfachansicht			
Funktionen	keine	Zoom	Drehung	Scrollen	Überblick/Detail
Komposition	Linear	Radial	Orthogonal	Frei	

Als Visualisierungsmethoden wurden u.a. folgende untersucht und klassifiziert: Sankey-Diagramm, Bubble-Chart, Collapsible-Treeview, Flare-Chart, Hierarchie-Bar, Chord-Chart, Sunburst-Chart, Heat-Map und Tree-Map.

4.3 Marktanalyse von NoSQL Datenbanken

Die Marktanalyse von NoSQL Datenbanken zeigte, dass neben den Open-Source-Variante hochinnovativer kleiner Anbieter mittlerweile auch große Anbieter den Nutzen erkannt haben, weshalb je Anwendungsfall spezielle Datenbanktypen am Markt vorhanden sind, die zu folgenden Klassen zusammengefasst werden können [10]:

- Key-Value-Datenbanken speichern Daten in Form von Schlüssel-Wertepaaren. Die Datenspeicherung erfolgt schemalos und ermöglicht dynamische Schemata, sodass Felder auch zur Laufzeit geändert und neu angelegt werden können.
- Spaltenorientierte Datenbanken speichern die Aggregationsdaten zusätzlich in Spalten.
- Dokumentenorientierte Datenbanken speichern ihre Daten schemalos in Dokumenten, die als Sammelmappe für Felder und Werte dienen. Sie sind für große Textmengen mit unbestimmter Länge einzusetzen.
- Graphen-Datenbanken bilden Beziehungen der Daten untereinander ab.

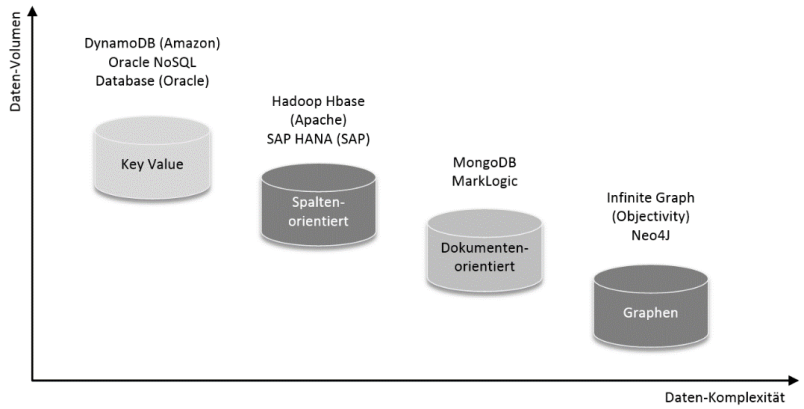

Abbildung 2. Übersicht NoSQL Datenbanken

Zum Aufbau eines geeigneten Prototyps wurde Hadoop gewählt. Hadoop ist ein Apache Software Foundation Open Source Entwicklungsprojekt und setzt sich zusammen aus dem Filesystem HDFS (Hadoop Distributed File System), der Programmierumgebung MapReduce und dem skalierbaren Datenhaltungssystem HBase.

Mittels dem MapReduce Programmiermodell können große unstrukturierte, sowie semistrukturierte Datensätze verarbeitet werden. MapReduce speichert die Daten in Blöcken und unterstützt auch die verteilte Speicherung auf mehrere Rechnereinheiten. Bei großen Datensätzen kann durch parallele Verarbeitung die Aufbereitungszeit reduziert werden. MapReduce bietet ein Framework zur Anwendungserstellung an sodass kein Code geändert werden muss. Das Konzept besteht aus den getrennten Abläufen Reduce und Map.

Die Verarbeitung beginnt mit dem Einlesen der unstrukturierten oder semistrukturierten Daten. Nach dem Einlesen erfolgt die Zuordnung des Inhalts der Dateien zu Positionen, wobei jede Zeile anhand eines Byte-Offsets identifiziert wird. Diese Schlüssel-/Werte-Paare werden in intermediate Schlüssel-/Werte-Paare transformiert. Darauf aufbauend werden gruppierte Schlüssel-/Werte Paare erzeugt, je Schlüssel werden die Werte zusammengefasst, sodass nur noch ein Wert pro Schlüssel vorhanden ist.

4.4 Prototyping

Anhand des Webshops und Webauftritts eines Unternehmens wurden die Daten aufbereitet und die Darstellungsmethoden gemäß erhobenen Klassifizierungen untersucht. Z.B. basierend auf „Visual Information-Seeking Mantra" [11] wurde die Einteilung und Bewertung der Informationsvisualisierung nach dem Prinzip „Overview first, zoom and filter, then details-on-demand" erstell. Gemäß dieser Richtlinie soll die Visualisierungsmethode zu Beginn einen Überblick über die Informationen darstellen. Darauf folgend sollen Filter- und Zoomfunktionen ermöglicht werden und zum Schluss einen detaillierten Einblick durch die Anwendung einer speziellen Selektierung geben.

Angelehnt an Shneiderman [11] wurden die beschrieben Datentypen mit den zugrunde liegenden Attributen der Elemente als Klassifizierungsgrundlage verwendet. Die weitere Selektion nach bestimmten Attributwerten erlaubt eine Einordnung der vorhandenen Daten. Wird von einer untersuchenden Visualisierungsmethode ein Datentyp nicht unterstützt, kann die Methode zur Darstellung ausgeschlossen werden.

Eine weitere Untersuchung erfolgte nach Freitas [12] mit der Unterscheidung der Visualisierungstechniken in zwei Gruppen. Einerseits die traditionellen Methoden zur Anzeige von Daten, Merkmale und Werte (Displays, Pseudo- Farbe, Höhenlinien und Vektorkarten zur Darstellung von Geo-Daten und unternehmensrelevanten Daten), sowie Techniken zur Anzeige von Datenstrukturen und Beziehungen (Informationsvisualisierung).

Als Ergebnis konnten Zuordnungen abgeleitet werden, welche Visualisierungsmethoden für welche Daten (strukturiert, sowie unstrukturiert) aus welchen Quellen geeignet sind.

Tabelle 2. Visualisierungsmethoden je Datenquelle

Datenquelle	Visualierungsmethode	Art der Daten
Kommunikationslog	Flare-Chart	Zusammenarbeit
Kommunikationslog	Chord-Chart	Zusammenarbeit (gewichtet)
Weblog	Calender-Chart	Anzahl Besucher
Weblog	Map	Geographische Darstellung
Weblog	Sankey-Diagramm	Aufrufe vor und nach einem Event
Webshop-Umsätze	Heat-Map	Absolut Umsätze oder Veränderungen
Webshop-Umsätze	Säulendiagramm	Produkt-Performane
Webshop-Umsätze	Bubble Chart	Umsatz und Menge
Webshop-Umsätze	Polar Bubble Chart	Umsatz, Menge und %-Anteil
Website-Clickstream	Map	Geographische Darstellung
Website-Clickstream	Säulendiagramm	Absprungraten
Tweets	Trendmap (Dienst)	Echtzeit Tweets

Auf Grund der gewählten Datenbasis, kann das Ergebnis für Web Analysen verwendet werden. In weiterer Folge sollen weitere unternehmensrelevante Daten z.B. aus dem Vertrieb, Produktion, Einkauf, usw. identifiziert und zusätzlich zu den bisherigen Auswertungen von strukturierten Daten mit allen prozessrelevanten Daten erweitert werden, um durch passende visuelle Methoden eine umfassendere Sicht auf die Prozesse im Unternehmen zu ermöglichen.

LITERATURVERWEISE

[1] Doug Laney, Information Economics, Big Data and the Art of the Possible with Analytics, Gartner, 2012

[2] Dwaine Snow, "Adding a 4th V to BIG Data - Veracity." Blog: Dwaine Snow's Thoughts on Databases and Data" Management, 2012
http://dsnowondb2.blogspot.com/2012/07/adding-4th-v-to-big-data-veracity.html

[3] Noel Yuhanna, The Forrester Wave: NoSQL Key-Value Databases, Forrester, 2014

[4] Analytics: Big Data in der Praxis, Wie innovative Unternehmen ihre Datenbestände effektiv nutzen Abbildung 6: Unternehmen verwenden hauptsächlich interne Datenquellen für Big Data-Initiativen. http://www-935.ibm.com/services/de/gbs/thoughtleadership/GBE03519-DEDE-00.pdf
[5] Judit Barnola, Ulrich Bäumer, Ulrich Baumgartner, Marialaura Boni, Konstantin Ewald, Angus Finnegan. The data gold rush, Growing and protecting your position in the data ecosystem. http://www.osborneclarke.com/media/filer_public/42/49/424982a2-09ad-4f70-b920-9bd50355984e/oc_digital_business_data_gold_rush_final.pdf
[6] Janna Quitney, Lee Rainie, Big Data: Experts say new forms of information analysis will help people be more nimble and adaptive, but worry over humans' capacity to understand and use these new tools well http://www.pewinternet.org/files/old-media/Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf
[7] Danah Boyd, Kate Crawford, Six Provocations for Big Data, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431
[8] Emmanuel Letouzé, Big Data for Development: Challenges & Opportunities, http://unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf
[9] Björn Bloching, Lars Luck, Thomas Ramge, Data Unser, Wie Kundendaten die Wirtschaft revolutionieren, Redline-verlag, 2012
[10] Paul Zikopoulos, Chris Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2011
[11] Ben Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, University of Maryland, 1996
[12] Carla M. Dal Sasso Freitas, Paulo R. G. Luzzardi, Ricardo A. Cava, Marco A. A. Winckler3, Marcelo S. Pimenta, Luciana P. Nedel, Evaluating Usability of Information Visualization Techniques, 2001