Seit einigen Monaten wird „Big Data“ intensiv, aber auch kontrovers diskutiert. Dieser Artikel zeigt nach einem einführenden Überblick anhand von Anwendungsfällen auf, wo die geschäftlichen Mehrwerte von Big-Data-Projekten liegen und wie diese neuen Erkenntnisse in die bestehenden Data-Warehouse- und Business-Intelligence-Projekte integriert werden können.

Analytische Mehrwerte von Big Data

Oliver Röniger und Harald Erb, ORACLE Deutschland B.V. & Co. KG

Der McKinsey-Report „Big Data“ betont die enorme gesellschaftliche und geschäftliche Bedeutung, die sich aus den explodierenden Datenmengen in nahezu allen Branchen ergibt [1]. Um tatsächlich von „Big Data“ zu sprechen, sind drei Merkmale zu erfüllen („3 Vs“):

- *Volume*
  Riesige Datenmengen (xx Terabyte), die sich bislang nicht für Data-Warehouse-Analysen erschließen lassen, weil deren relevante Informationsdichte einfach zu gering ist, als das sich deren Speicherung und Verarbeitung aus wirtschaftlicher Sicht lohnt.
- *Velocity*
  Die hektische zeitliche Frequenz, in der Daten in operativen Geschäfts-Prozessen entstehen. Mehrwerte werden sowohl aufgrund der sehr hohen Granularität der Daten als auch in deren umgehender Verarbeitung und Erkenntnisgewinnung in Echtzeit gesehen.
- *Variety*
  Die Vielfalt der zusätzlichen (unstrukturierten) Datenformate, die sich jenseits der üblichen wohlstrukturierten Transaktionsdaten aus Social-Media-Daten, Maschine-zu-Maschine-Kommunikationsdaten, Sensordaten, Webserver-Logdateien etc. ergeben.

Diese Daten sind inhaltlich neu, sie sind unstrukturiert, es sind unsagbar viele – die wirklich interessanten Informationen darin sind hingegen nur äußerst dünn gesät. Insofern liegt es nahe, sich an das folgende einfache Vorgehensmodell zu halten:

1. Gezieltes Sammeln der neuartigen Massendaten aus den relevanten Datenquellen
2. Filtern dieser Daten aufgrund definierter interessanter Merkmale
3. Selektive Weiterverarbeitung beziehungsweise Übernahme der interessanten Informationen in die vorhandenen internen IT-Systeme
4. Die verarbeiteten Daten aus dem 1. Schritt wegwerfen und den Prozess fortsetzen

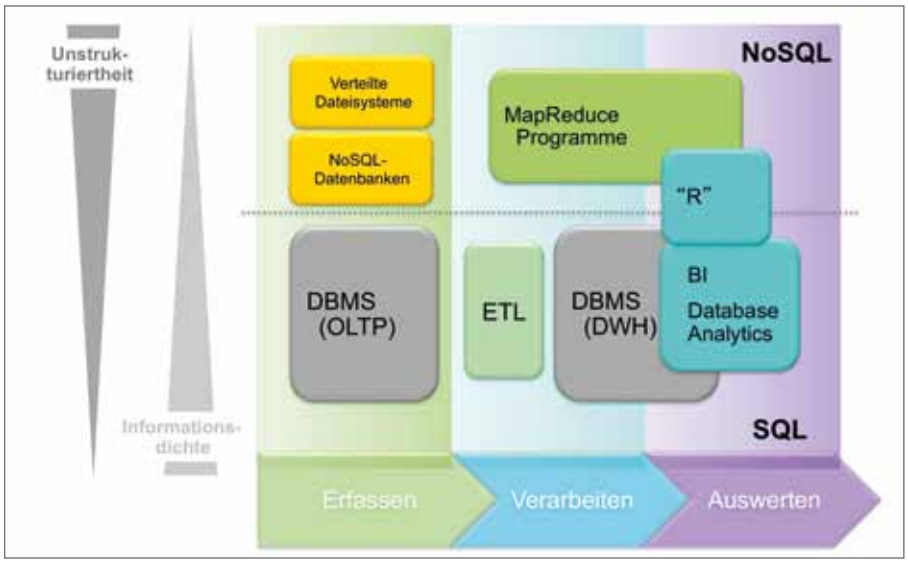

Abbildung 1: Gegenüberstellung der Komponenten

Um diese unstrukturierten, schemalosen Daten überhaupt sammeln zu können, wurden von Google und anderen Internet-Pionieren NoSQL-Datenbanken (wie Cassandra) entwickelt und mit Hadoop sowohl ein verteiltes Dateisystem (HDFS) als auch ein Entwicklungs-Framework (MapReduce) bereitgestellt (siehe Positionierung der Oracle Big Data Appliance [2]). Abbildung 1 stellt die maßgeblichen Komponenten der NoSQL- und SQL-Welt gegenüber.

Zunächst soll eine mögliche gemeinsame Architektur betrachtet werden, um diese Technologien parallel oder auch gemeinsam zu betreiben, bevor aus Anwendungssicht die Frage geklärt wird, was dieses pragmatische Vorgehensmodell konkret für verschiedene Anwendungsfälle bedeutet.

Zusammenspiel Big Data/Data Warehouse

Bei einer klassischen Konzeption eines Data-Warehouse und Business-Intelligence-Systems, leicht modifiziert nach [3], bleiben durch Big Data die bestehenden Data-Warehouse- und Business-Intelligence-Prozesse zunächst unangetastet. Die neuartigen Datenquellen erweitern aber zum einen den analyserelevante Datenraum, was Erkenntnisgewinn verspricht, zum anderen treten an die Seite von klassischen BI-

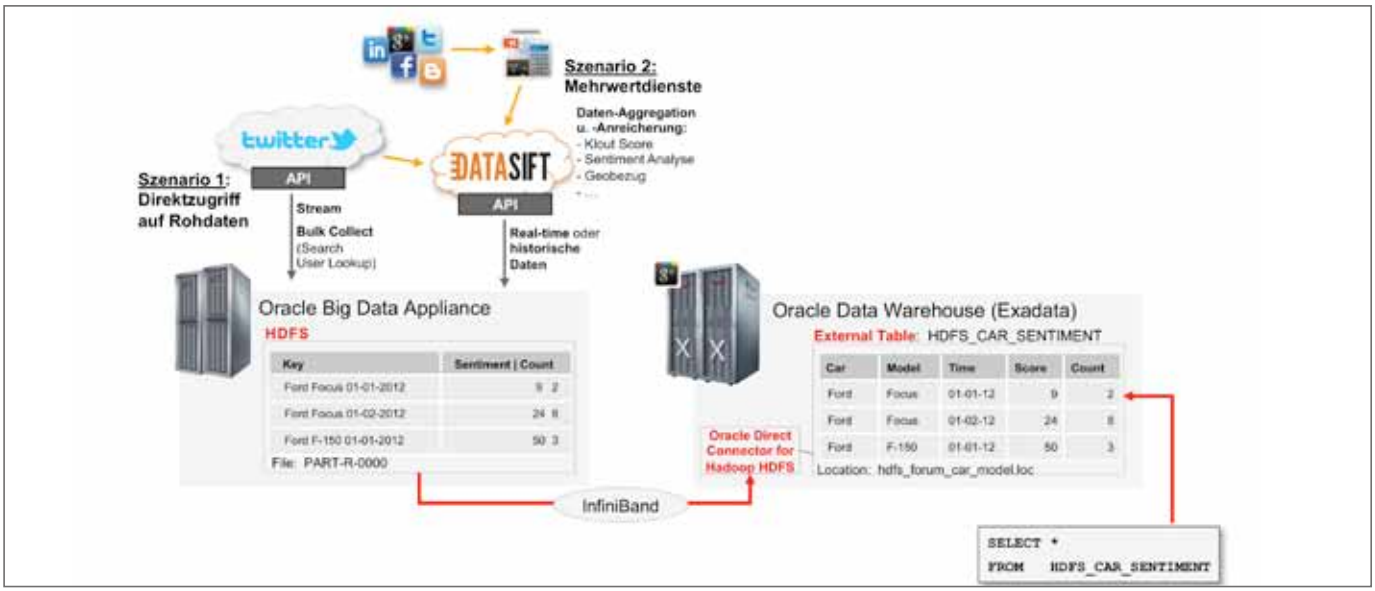

Abbildung 2: Vom Twitter-Feed zum Big-Data-Zugriff via External Table im Data Warehouse

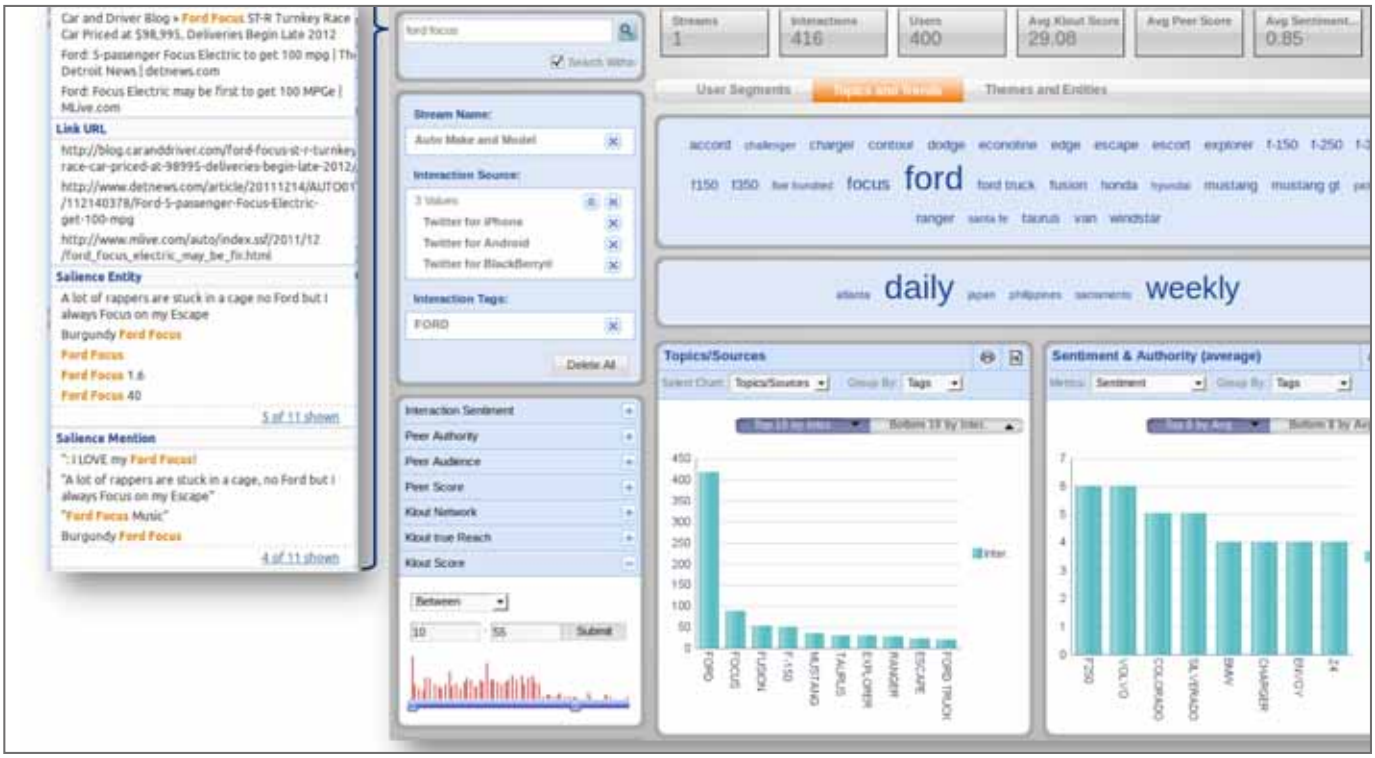

Abbildung 3: Beispiel eines Endeca-Dashboards

Werkzeugen zusätzliche Suchfunktionalitäten, die den unstrukturierten, textuellen Informationen besser gerecht werden. Es handelt sich jeweils um Ergänzungen zum Bestehenden, also eher Evolution als Revolution. Eine technische Kernfrage lautet, wie die unstrukturierten Massendaten aus Big Data mit dem Data Warehouse verbunden werden können. Hierzu gibt es seitens Oracle mehrere technische Möglichkeiten:

- *Oracle Loader for Hadoop*
  Daten aus einem Hadoop-Cluster werden direkt in das Oracle Data Warehouse geladen
- *Oracle Direct Connector for Hadoop HDFS*
  Direkter Zugriff auf das verteilte Filesystem für das Oracle Data Warehouse
- *Oracle Data Integrator (ODI) Application Adapter for Hadoop*
  Einbinden eines Hadoop-Jobs in einen ODI-Ladeprozess

Abbildung 2 zeigt beispielhaft anhand von Twitter-Nachrichten zwei unterschiedliche Szenarien, wie sogenann-

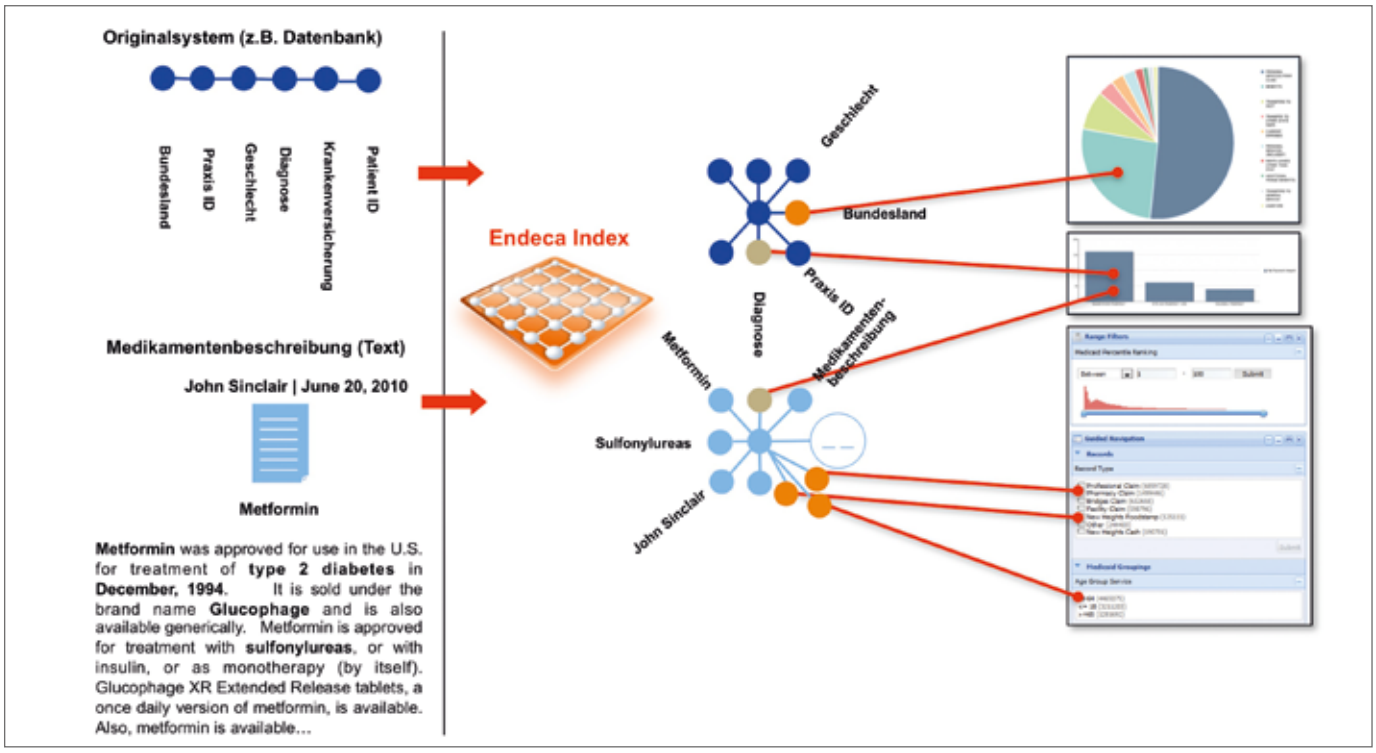

Abbildung 4: Beispiel eines Facetten-Datenmodells

te „Social-Media-Daten" in die Big-Data-Infrastruktur eines Unternehmens überführt und auswertbar gemacht werden können. Szenario 1 steht dabei für den individuellen Entwicklungsansatz, bei dem die Akquisition der Rohdaten über Twitter-Developer-APIs (siehe http://dev.twitter.com) und die Datenorganisation über das Hadoop-MapReduce-Entwicklungs-Framework (nicht abgebildet) erfolgt. Alternativ lassen sich heute auch schon Mehrwertdienste (Szenario 2) in Anspruch nehmen, die per Auftrag Twitter-Datenabzüge aufbereiten und anreichern, indem sie unter anderem den Geo-Bezug herstellen, den Einfluss der Twitter-Beiträge auf andere per „Klout Score" ermitteln oder eine Sentiment-Analyse durchführen. Im Ergebnis werden die relevanten Daten (in der Abbildung die

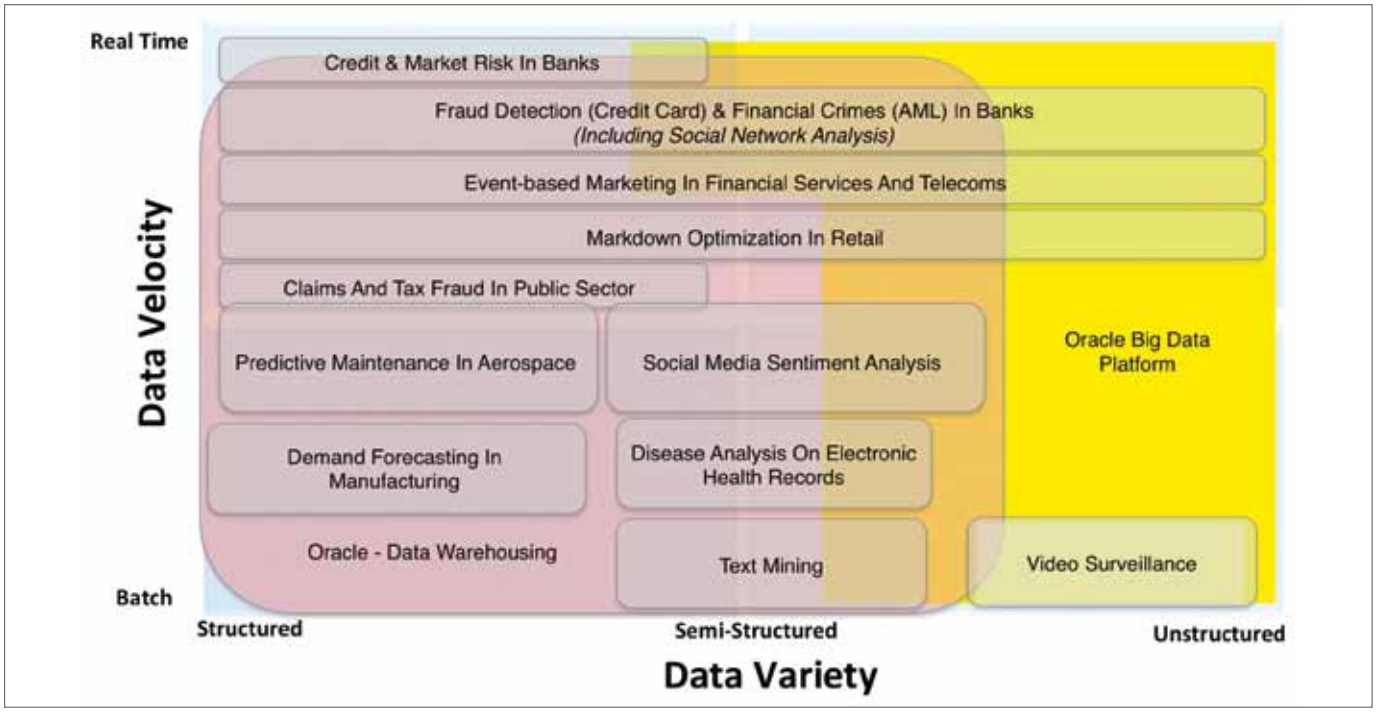

Abbildung 5: Big-Data-Anwendungsbereiche: Oracle Lösungsquadrant

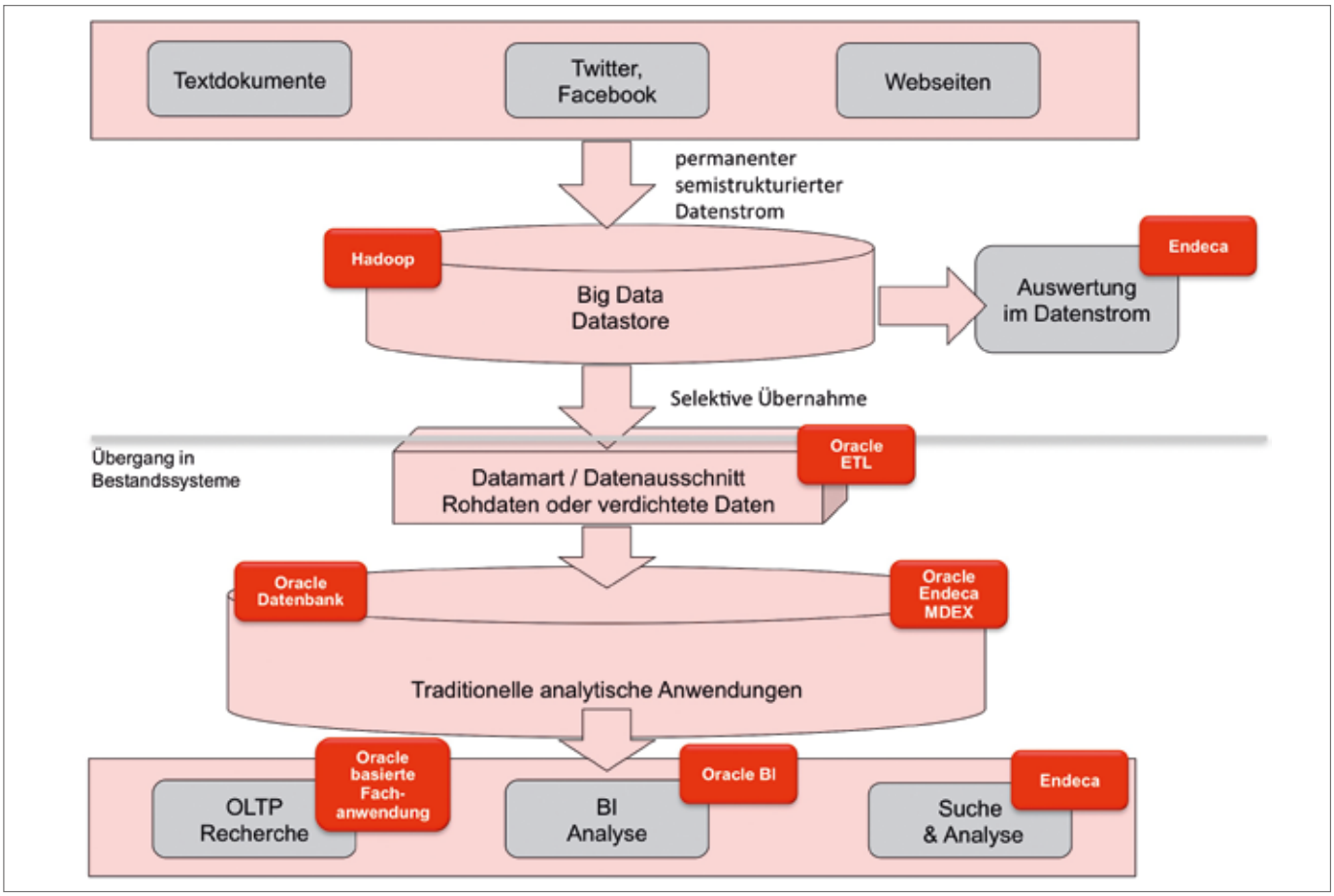

Abbildung 6: Analytisches Gesamtszenario

veredelten Twitter-Feeds) als Key-Value-Paare in Dateiform in einem Hadoop Distributed File System (HDFS) zur weiteren Analyse bereitgestellt. Nutzt man hierzu die Oracle-Big-Data-Infrastruktur in Kombination mit einem Oracle-Data-Warehouse, eröffnet sich dem Analysten ein eleganter Weg des Datenzugriffs per External Tables und SQL (siehe auch [4]).

Analysemöglichkeiten

Die Akquisition von Endeca erweitert das bisherige Oracle-Business-Intelligence-Analyse-Spektrum, indem die textbasierte Suche unstrukturierter Informationen mit den typischen quantitativen BI-Analysen kombiniert und dem Benutzer intuitiv nutzbar präsentiert wird. Die Verbindung quantitativer und qualitativer Informationen überschreitet die klassische Grenze von Business Intelligence und kann konzeptionell dem Knowledge Management zugeordnet werden. Der Slogan „No data left behind“ drückt diese Philosophie treffend aus. Erweiterte Analyse-Funktionen sind zum Beispiel die unternehmensweite Suche, die Präsentation in Form von Tag Clouds, das datengetriebene dynamische Filtern von Merkmalen und die sogenannte „Facetten-Navigation“, bei der die Suche und Auswahl von Attributen wie auf einer Webseite funktioniert [5].

Abbildung 3 zeigt plastisch Teile dieser neuen funktionalen Möglichkeiten. Es geht um die Analyse eines Twitter-Streams zum Thema „Auto Make and Model“. In der Guided-Navigation-Leiste links sieht man die einbezogenen Datenquellen (iPhone-, Android- und Blackberry-Nutzer) und die weiteren gesetzten Filterkriterien („Ford Focus“). Oben in der Metrik-Leiste wird ausgewiesen, dass in 416 (von ca. 350.000 Interaktionen) zutreffende Nachrichten gefunden wurden und sich 400 (der ca. 132.500 Benutzer) zu diesem Thema austauschen. In den Tag-Clouds werden besonders häufig verwendete, unterschiedliche Pkw-Modelle und andere Begriffe hervorgehoben, wobei die Größe der Schrift zeigt, auf welche Wörter die meisten Treffer kommen. Die bereits erwähnten Möglichkeiten zur Anreicherung von Social-Media-Daten durch „Klout Scores“ und Sentiment-Analysen helfen dem Analysten bei der Bewertung der Twitter-Beiträge, etwa in Form zusätzlicher Metriken oder weiterer Attribute für die geführte Suche im Datenbestand. Schließlich finden sich unten weitere Statistiken, die zusätzlichen korrespondierenden Inhalt enthalten können.

Bevor es zur fachlichen Analyse kommen kann, sind die Daten aufzubereiten, gegebenenfalls zu verknüpfen sowie anzureichern. Neben klassischen ETL-Funktionen gibt es seitens Endeca ein erweiterbares Content-Acquisition-System (CAS) für die Daten-Integration von Hunderten von Dateitypen, Dokument-Repositories, CMS-Systemen, Webinhalten und RSS-Feeds. CAS kann sowohl Dateiserver als auch Twitter, Facebook & Co. ana-

lysieren. Jedes unstrukturierte Attribut kann verarbeitet und um weitere Informationen angereichert werden. Gängige Techniken sind:

- Automatic Tagging
- Named Entity Extraction
- Sentiment Analysis
- Term Extraction
- Geospatial Matching

Die unstrukturierten Daten können mit anderen Datensätzen über einen beliebigen Schlüssel miteinander verbunden werden. Natürlich können auch strukturierte Daten mit diesen unstrukturierten Daten im Rahmen des ETL-Prozesses verknüpft sein. Dabei wird keine feste analysefokussierte Datenmodellierung betrieben – wie im Data Warehouse in Richtung Star- oder Snowflake-Modell in Form von fest verknüpften Tabellen üblich –, sondern die Dimensionen werden alle gleichberechtigt nebeneinander in ein Modell gelegt. In der Praxis existieren Analyse-Modelle mit mehreren Hundert Dimensionen. Aus fachlicher Sicht eröffnen sich so unendliche Analyse-Möglichkeiten. Abbildung 4 veranschaulicht die Idee des hochdimensionalen Facetten-Datenmodells.

Die Praxis

Big-Data-Projekte sind kein Selbstzweck. Die neue Technik ist reizvoll, aufgrund des notwendigen Spezialwissens und der sehr großen Datenmengen (Hardware-Bedarf) aber durchaus kostenintensiv. Daher ist es erforderlich, die fachlichen neuen Möglichkeiten, die sich aus Big-Data-Analysen ergeben können, nüchtern zu bewerten. Das kann nur jedes Unternehmen selbst anhand seiner Anwendungsfälle tun. In Anlehnung an [6] zeigt Abbildung 5 eine Gegenüberstellung einiger Big-Data-Anwendungsbereiche und des Oracle-Lösungsangebots zu Big Data und Data Warehousing.

Unter www.doag.org/go/doagnews/erb_tabelle sind beispielhaft fünf ausgewählte Use Cases vorgestellt und ihre Komplexität sowie deren Geschäftsnutzen bewertet.

Quellenverzeichnis

[1] McKinsey Global Institute: Big Data: The next frontier for innovation, competition, and productivity, Report, May 2011: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

[2] Carsten Czarski, Big Data: Eine Einführung, Oracle Dojo Nr. 2, München 2012: http://www.oracle.com/webfolder/technetwork/de/community/dojo/index.html

[3] Cackett, D./Bond, A./Lancaster,K./Leiker, K., Enabling Pervasive BI through a Practical Data Warehouse Reference Architecture, An Oracle White Paper, Februar 2010: http://www.oracle.com/us/solutions/datawarehousing/058925.pdf

[4] Günther Stürner, Big Data – Hype und Wirklichkeit, Vortrag auf dem Führungskräfte-Forum „Ergebnis- und wirkungsorientierte Steuerung“ des Behördenspiegels: http://www.fuehrungskraefte-forum.de/?page_id=1617

[5] Mark Rittman, Where Does Endeca Fit with Oracle BI and DW?, 22. Februar 2012, http://www.rittmanmead.com/2012/02/oracle-endeca-week-where-does-endeca-fit-with-oracle-bi-dw-and-epm/

[6] Ravi Kalakota, Big Data Analytics Use Cases, 12. Dezember 2011: http://practicalanalytics.wordpress.com/2011/12/12/big-data-analytics-use-cases

[7] TU München, o.V., Neuer Krebsauslöser in Pommes frites entdeckt; scinexx – Das Wissensmagazin, 19. August 2008, http://www.g-o.de/wissen-aktuell-8686-2008-08-19.html

[8] o.V.: Bei Twitter hat Obama im Wahlkampf die Nase vorn, in Westdeutsche Allgemeine Zeitung Online, 3. Januar 2012, http://www.derwesten.de/wirtschaft/digital/bei-twitter-hat-obama-im-wahlkampf-die-nase-vorn-id6210915.html

[9] o.V.: Neue Umsatzsteuer soll Betrug vorbeugen, in Frankfurter Allgemeine Zeitung Online, 20. Oktober 2005: http://www.faz.net/aktuell/wirtschaft/wirtschaftspolitik/haushalt-neue-umsatzsteuer-soll-betrug-vorbeugen-1282102.html

Oliver Röniger
oliver.roeniger@oracle.com

Harald Erb
harald.erb@oracle.com

Wir begrüßen unsere neuen Mitglieder

Persönliche Mitglieder

Norbert Kossok
Dirk Wemhöner
Alexandra Strauß
Thomas Ewald-Nifkiffa
Kevin Brych
Joachim Engel
Thorsten Grebe
Martin Bernemann
Josef Rabacher
Uwe Schreiber
Wolfgang Michael Girsch
Christa Weckman
Thomas Krahn
Marco Stroech
Wolfgang Bossmann
Christoph Mecker
Corinna Kerstan
Gerhard Schaefer
Michael Tucek
Rüdiger Ziegler
Erika Krüger
Andreas Koop
Ulrich Gerkmann-Bartels
Manfred Drozd
Christoph Quereser
Andreas Reinhardt
Markus Vincon

Firmenmitglieder

Dirk Fleischmann, cubus BI Solutions GmbH
Wolfgang Hack, dimensio Informatics GmbH
Volker Oboda, DMySQLAG e.V.
Martin Böddecker, mb Support GmbH
Hans Haselbeck, EMPIRIUS GmbH