Das Integrieren neuer Informationsquellen und ihrer Auswertungen gewinnt für Firmen immer mehr an Bedeutung. Dazu gehören auch Daten aus sozialen Netzwerken wie Twitter oder Facebook. Entsprechend steigen die Datenmengen in allen Unternehmensbereichen rasant an.

Social Data Analyse

Norbert Henz, Trivadis GmbH

Die Herausforderung besteht darin, aus dieser Vielfalt an Datenmaterial relevante Informationen herauszufiltern. Es gilt, schnell und einfach neue Erkenntnisse aus einer flexiblen Verknüpfung unterschiedlichster Datenquellen zu gewinnen. Solche Systeme müssen schnell anpassbar, dabei variabel und leicht zu bedienen sein. Der heutige Anwender will nicht mehr monatelang auf seinen Datenzugriff warten.

Mit Oracle Endeca Information Discovery können Daten aus unterschiedlichsten Quellen schnell und einfach zueinander in Bezug gesetzt und dem Anwender umgehend zu Analyse-Zwecken angezeigt werden. Mit seinen „Search and guided Navigation"-Features erlaubt diese Lösung schnelle Antwortzeiten und gleichzeitig die freie Auswahl von Such-Optionen für die Endanwender.

Durch die hohe Flexibilität bei der Daten-Zusammenstellung lassen sich neue Daten zügig zu bestehenden Datastores hinzufügen und stehen dem Anwender somit zeitnah zur Verfügung. Der Artikel stellt anhand eines Lösungsbeispiels die Möglichkeiten von Endeca vor.

Die Ausgangssituation

Die Firma Trivadis sucht immer nach guten, qualifizierten Beratern und stellt entsprechende Anstrengungen an, um geeignete Personen auf sich aufmerksam zu machen. Daher werden die Stellenbeschreibungen nicht nur auf der Homepage veröffentlicht, sondern schon lange zusätzlich unter anderem auch via Twitter beworben.

Als soziales Netzwerk ist Twitter ein ideales Medium, um in direkteren Kontakt mit anderen Menschen zu treten. Mittels der Hashtags, das sind die Textteile mit dem # davor, werden die Meldungen („Tweets" genannt) zusätzlich kategorisiert.

Trivadis verschickt bei Twitter seine Stellenangebote mit dem Hashtag #Jobs. Wer dem Twitter-Stream folgt, wird über unsere Stellenangebote auf diese Art informiert. Aber auch jeder andere Mensch, der bei Twitter nach Jobs sucht, erhält die Anzeigen. Einfach, effektiv und mit wenig Aufwand wird durch Twitter die Reichweite der Stellenanzeigen erhöht.

Die Herausforderung

Aber erreicht man mit dieser Maßnahme auch wirklich jemanden? Wird via Twitter auf die Stellenanzeigen zugegriffen? Wenn ja, wie oft? Um solche Fragen beantworten zu können, muss man die Informationen von Twitter mit den Stellenanzeigen verknüpfen und aus dem Twitter-Stream die relevanten Informationen gewinnen. Nun hat der Twitter-Stream zwar eine Struktur, aber viele interessante Detail-Informationen liegen lediglich in Textform vor. Endeca bietet die Funktionalität, solche unstrukturierten Daten mit strukturierten Daten in einem gemeinsamen Datastore zu verbinden und somit analysierbar zu machen.

Mit der Definition eines Datastores hält der Endeca Server die geladenen Daten im Hauptspeicher. Es kommt also eine In-Memory-Lösung zum Einsatz. Durch diese Form der Datenhaltung im Hauptspeicher und seine sehr flexible Datenstruktur erlaubt Endeca eine schnelle Anpassung an sich ändernde Berichtsanforderungen. Die Entwicklungszyklen können hierbei sehr kurz gehalten sein. Erste Daten laden, ad-hoc auswerten, die Ladestrecke wieder anpassen und erneut analysieren. Das ist schnell machbar und flexibel anpassbar. Das Endeca-ETL-Werkzeug „Integrator" öffnet

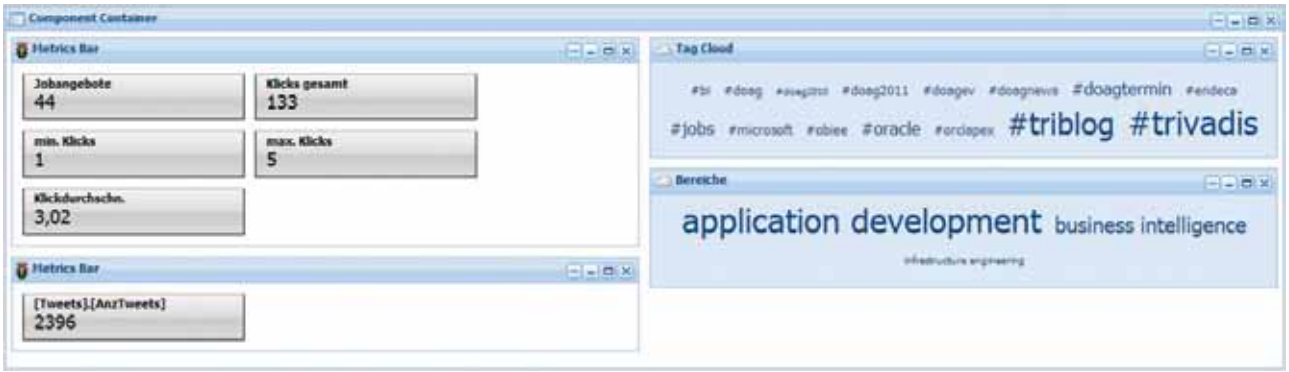

Abbildung 1: Preview auf Twitter-Daten mit Hashtag

die Datenquellen und extrahiert den gewünschten Inhalt. Dahinter versteckt sich eine erweiterte Version des um viele Funktionalitäten ergänzten Open-Source-ETL-Werkzeugs „CloverETL“.

Nach Umformung und Aufbereitung im ETL-Prozess werden die Daten dann in einem Datastore abgelegt. Für die Daten-Transformation liefert Endeca eine umfangreiche Bibliothek an nützlichen Funktionen im Endeca Integrator mit (siehe Abbildung 2).

Die Verknüpfung der Daten aus den verschiedenen Quellen erfolgt ganz simpel über gemeinsame, gleichartige Attribute. Diese ermöglichen dem Anwender bei seinen Auswertungen die übergreifenden Abfragen auf alle geladenen Daten in diesem Datastore. Durch die In-Memory-Technik bietet der Datastore dabei eine hohe Flexibilität und ermöglicht sehr schnelle Abfragen auch bei großen Datenmengen.

Sollten die Quell-Datensätze einmal nicht über die notwendigen Attribute für eine Verknüpfung verfügen, muss der Entwickler eingreifen. Durch Extraktion aus vorhandenen Strings kann er zum Beispiel Teil-Inhalte herauslösen und diese als neues, gemeinsames Attribut zusätzlich in den Datastore laden.

Da die Daten ihre ursprüngliche Daten-Strukturen und -Typen beibehalten, ergibt sich eine sehr unterschiedliche Gesamt-Datenstruktur im Datastore.

Erkennbar ist, dass es keine vordefinierten Tabellen gibt; die Datenstruktur entsteht durch die Daten aus den verschiedenen Quellen von selbst. Die Datensätze einer Quelle müssen dabei noch nicht einmal die gleiche Satzlänge haben. Einzig wichtig für die geplanten Auswertungen sind die gemeinsamen Attribute, hier in der Mitte des Bildes im roten Bereich gezeigt. Aufgrund dieser Schlüssel-Attribute kann später quellübergreifend ausgewertet werden. Ohne diese Verbindungs-Attribute würden die einzelnen Inhalte bezugslos nebeneinander im Datastore liegen.

Ein Beispiel

Zu allererst müssen aus den Twitter-Daten Inhalte extrahiert werden, um als Schlüssel-Attribute oder Filter-Kriterium zur Verfügung zu stehen. Über die Hashtags ist es für den Entwickler sehr einfach, dies zu bewerkstelligen. Die Hashtags verbergen sich im eigentlichen Text einer Twitter-Meldung. Durch einen regulären Ausdruck erlaubt Endeca, diese zu erkennen.

Der Operator „TEXT_TAGGER_REGEX“ im Importer filtert über den regulären Ausdruck die Hashtags heraus und schreibt sie nachfolgend in ein neu definiertes Attribut. In der Vorschau auf die Twitter-Daten sind die Hashtags im Textfeld sehr gut zu erkennen (siehe Abbildung 1).

Nach der Anwendung des Operators sieht man das Ergebnis: Die Hashtags wurden erkannt und separat gespeichert. Damit stehen diese Inhalte jetzt für Auswertungen zur Verfügung. Dies ist nur ein kleines Beispiel für die vielfältigen Möglichkeiten von Endeca Integrator, um Daten aufzubereiten.

Datenauswertung mit Oracle Endeca

Die Auswertungen selbst erfolgen per Web-Browser durch Endeca Information Discovery. In dieser Anwendung kann sich der Anwender flexibel, schnell und interaktiv in den Daten des Datastores bewegen (siehe Abbildung 2). Anpassungen an den Dashboards sind jederzeit selbstständig realisierbar. Durch die geführte Suche in der bereit-

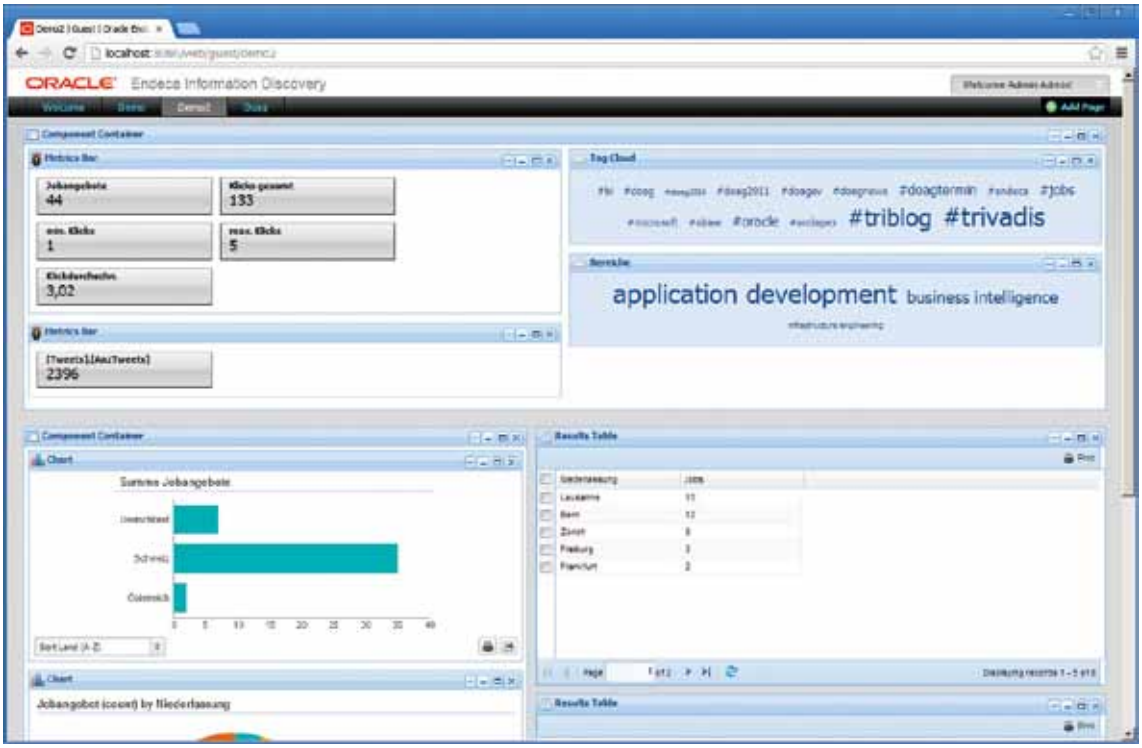

Abbildung 2: Beispiel eines Endeca Dashboards

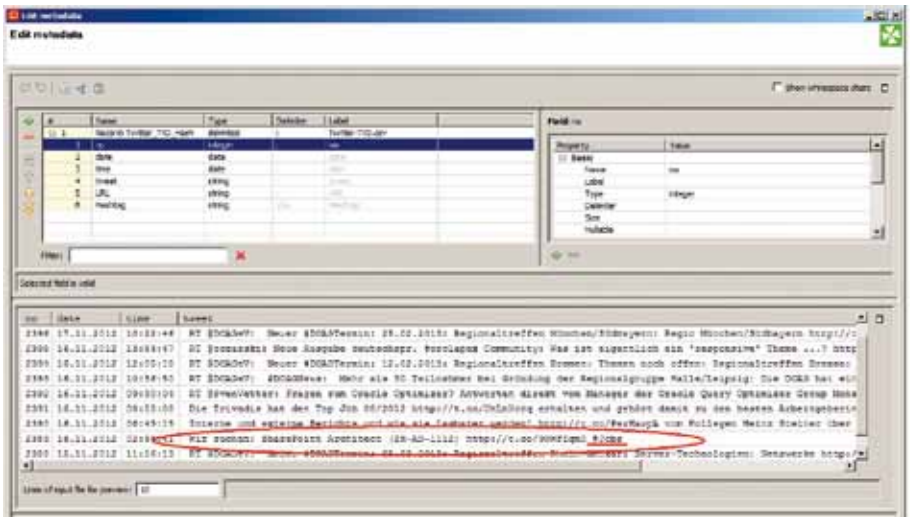

Abbildung 3: Teil einer Auswertung mit Tag-Wolke und Metrik-Auswertung

gestellten Datenwelt bietet Endeca Discovery mehr als nur ein vorgefertigtes Reporting oder eine Analyse. Der Anwender selbst bestimmt die Richtung seiner Auswertungen, er erforscht quasi seine Datenwelt und kann so zu völlig neuen Erkenntnissen kommen.

Für die Trivadis-Stellenanzeigen wurden einige verschiedene Möglichkeiten der Auswertung getestet. So kann man sehr einfach die Anzahl der Anzeigen zählen, aber auch die Zahl der Klicks darauf ermitteln, was schon interessanter ist. Über die Verteilung auf die Fachbereiche erkennt man dann schnell, welche Angebote am meisten Interesse geweckt haben (siehe Abbildung 3).

Zudem lassen sich die Inhalte durch grafische Darstellungen visuell präsentieren. Sehr schön ist die Interaktivität aller Darstellungskomponenten gelungen. Der Anwender wandert quasi durch die neuen Daten und gibt seiner Analyse selbst die gewünschte Richtung.

Von übergeordneten Sicht-Ebenen sind die Details aller Daten immer erreichbar. Auch kann durch Verlinkung jederzeit auf andere Inhalte verwiesen werden, beispielsweise auf die Original-Stellenanzeige.

Fazit

Die Auswertung der Twitter-Daten hat gezeigt, dass die Stellenangebote über diesen Weg sehr wohl Beachtung finden. Daraus haben sich interessante Einblicke und Handlungs-Optionen ergeben.

In zukünftigen Schritten lässt sich auf dieser Basis die Twitter-Auswertung auf weitere Themen ausdehnen. Zusätzliche Datenquellen können zum Datastore hinzugefügt werden und ermöglichen so eine noch tiefere Analyse zusammen mit unseren internen Daten.

Norbert Henz
norbert.henz@trivadis.com