Social Data Analyse mit Oracle Endeca

Norbert Henz
Trivadis GmbH
Hamburg

Schlüsselworte
Oracle Endeca, Socialdata, Search&Analyse, In-Memory-Technology

Einleitung
Oracle Endeca gehört erst seit kurzem zum Produktportfolio von Oracle. Diese Produkte ergänzen die bekannten Oracle BI-Tools der Oracle BI-Suite um eine Reihe neuer Möglichkeiten zur Analyse von strukturierten und unstrukturierten Daten.In diesem Vortrag werden die EndecaProdukte kurz vorgestellt und in einem praktischen Beispiel ihr Nutzen als BI-Tool demonstriert.

Vorgeschichte
Die Firma Endeca wurde bereits 1999 in Cambridge, MA gegründet. Im Oktober 2011 wurde die Firma dann von Oracle aufgekauft. Zu dieser Zeit hatte die Firma um die 500 Mitarbeiter.

Der Name Endeca leitet sich vom deutschen Wort ‚entdecken' ab. Und das beschreibt recht deutlich, wozu diese Software-Produkte erstellt wurden. Das Ziel war es, dem Endanwender eine einfache und schnelle Lösung zur Sichtung von Suchergebnissen im Internet zur Verfügung zu stellen. Der Aufwand in der Erstellung einer BI-Lösung und der dazu notwendige Lernaufwand sollten dabei möglichst klein gehalten werden. Durch die Kombination von InMemory Datenhaltung, in Form von Key-Value-Paaren mit einem analytischen Frontend kann der Anwender sehr schnell erste Ergebnisse erzielen. Die Entwicklung einer Endeca Anwendung ist vom Ansatz her iterativ und auf Veränderung ausgelegt. Bei den Filtern, der Navigation und der Suche kann jedes Attribut aus den Datensätzen verwendet werden. Dabei erfolgen die Berechnungen fortlaufend, so dass jede Änderung zu einer weiteren Sicht auf die Daten führt.

Oracle Endeca
Oracle hat nach dem Kauf die Endeca Produkte in sein Portfolio integriert und umbenannt:

- Latitude wurde zu Oracle Endeca Information Discovery Studio (Frontend)
- MDEX Engine wurde der Oracle Endeca Server (Backend)
- Integration Suite (ETL-Tool) wurde zum Oracle Endeca Information Discovery Integrator

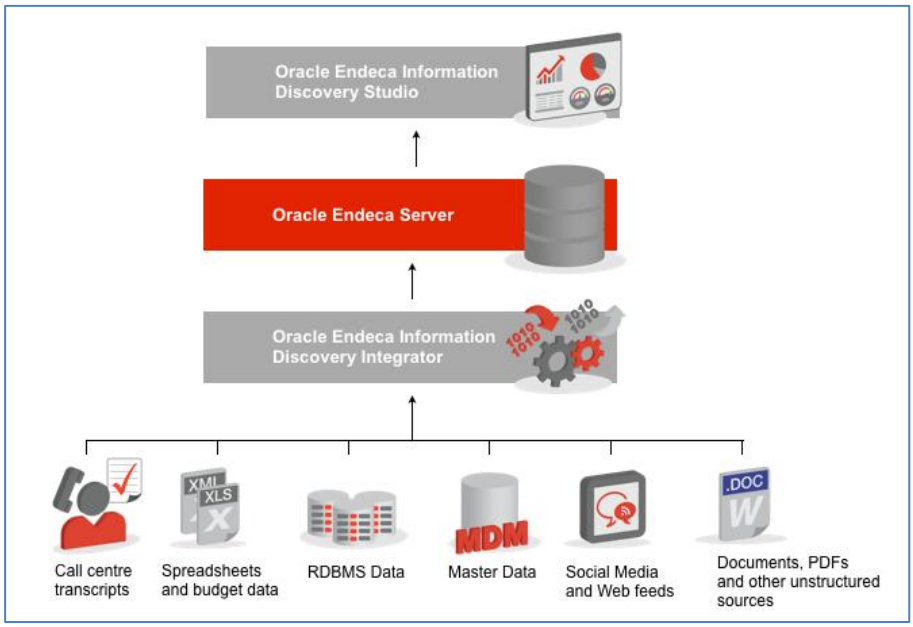

Abb. 1: Oracle EndecaProduktstack

Das Information Discovery Studio bietet eine grafische Oberfläche zur Datenanalyse. Sie arbeitet webbasiert. Mit dem Studio kann der Anwender sehr schnell, auch prototypisch, in die Analyse seiner Daten im Endeca Server einsteigen.

Der Endeca Server (MDEX) ist als spaltenorientierte InMemory-Lösung für die Datenhaltung entwickelt worden. Es handelt sich um keine Datenbank. Dieser Server arbeitet stateless, d.h. er merkt sich keine Sitzungsinformationen und vorherige Abfragen. Zusätzlich verfügt er über ganz spezielle eigene Datenstrukturen und Algorithmen, anders als eine relationale Datenbank.

Der Information Discovery Integrator steht als ETL Werkzeug für die Befüllung des Endeca Servers (MDEX) zur Verfügung. Sie basiert auf dem Open Source Produkt Clover-ETL und wurde um spezielle Funktionen für den Endeca Server ergänzt.

Zwischen all diesen Produkten erfolgt der Datenaustausch durch spezielle Endeca-Webservices. Diese Webservices existieren für die verschiedensten Funktionalitäten beim Daten laden sowie beim Zugriff auf die Daten im Server über das Studio. Zusätzlich gibt ein Bulk-Loader-Interface um große Datenmengen zu laden.

Die Endeca Information Discovery Hardware-Anforderungen sind:

Als 64-bit Applikation kann Endeca Multi-CPUs/Cores nutzen und viel (!) RAM zur Verfügung haben.
Oracle Information Discovery 2.3 gibt es als Version für Windows und Linux
Windows Server 2008 SP1 64-bit
Oracle Linux / RHEL 5 64-bit

Aufbau der Demo

Im Rahmen der Demo werden Daten aus Twitter und anderen Quellen mittels der Integrations Suite (ETL) in den Endeca Server geladen und dort verknüpft. Durch Hinzufügen von weiteren Informationen kann anschließend mit dem Studio eine Analyse und Suche in diesen Daten erfolgen.

Inhalte der Demonstation:
Im ersten Schritt werden die Daten geladen:

Der Endeca Server läuft als Serverprozess. Mit diesem Prozess wird der notwendige Datastore für die Datenhaltung im RAM definiert und anschließend zur Verfügung gestellt.
Die Erstellung eines Datastores kann mit der Endeca Kommandozeile „`cd c:\Oracle\Endeca\Server\2.3.0\endeca-cmdendeca-cmdcreate-ds<Name>`" oder mittels der grafischen Oberfläche in der Integration Suite erfolgen. Bei beiden Wegen wird auf die gleiche Webservice API zugegriffen.
Der neue Datastore wird nun mit den ersten Datensätzen (Records) durch die Integration Suite gefüllt. In dieser Demo kommen der Einfachheit halber alle Records aus einer Datei. Alle Datensätze müssen eine Spezifikation und einen Unique Key besitzen. So kann jeder dieser Datensätze eindeutig identifiziert und adressiert werden.
Nach dem Laden können diese ersten Datensätze bereits mit dem Studio anzeigt werden.
Nachfolgend werden weitere Datenquellen in bestehenden Ladegraphen eingefügt. Durch Joins werden diese neuen Daten mit den bereits vorhandenen ersten Datensätzen verknüpft. Dabei ist es egal, ob es sich um strukturierte oder unstrukturierte Daten handelt. Wichtig ist nur, dass es immer mindestens einen gemeinsamen Key zur Verknüpfung untereinander gibt. Nur so kann später das Studio alle Informationen verarbeiten.
Dieser Vorgang kann nun solange fortgeführt werden, bis alle gewünschten Daten auf diesem Wege in den gemeinsamen Datastore geladen wurden. Beim Laden können Textinhalte mit Endeca analysiert, angereichert und geparsed werden und so zusätzliche Darstellungsmöglichkeiten dem Studio zur Verfügung stellen.

Informationen analysieren und visualisieren:

Für den Anwender stellt das Studio den Zugang zu allen Daten in einem Datastore dar. Mit dem Studio kann er bereits nach jedem erfolgreichen Ladelauf auf den Inhalt im Datastore zugreifen.
Das Studio arbeitet webbasiert im Browser. Durch den frühzeitigen Zugang entwickelt sich vom ersten, prototypenhaft Schritten die eigentliche Analyseanwendung. Schritt für Schritt erstellt, testet und verbessert so der Anwender seine Applikation. Im Laufe dieses Entwicklungsprozesses entstehen so auch komplexe Anwendung mit unterschiedlichsten Datendarstellungen und Diagrammformen.
Meist beginnt die Entwicklung mit ersten Tabellen, dann kommen Suchfeldern für Werte oder Datensätze hinzu und weitere Datendarstellungen und Grafiken. Auch Geodarstellungen auf Kartenmaterial sind mit Endeca möglich. Das Studio bietet eine Vielzahl an Darstellungsoptionen für die verschiedensten Einsatzszenarien an.Bei all diesen Optionen kann das Layout flexibel angepasst und so ansprechende Dashboards für die Datenanalyse ermöglicht werden.

In der Demo werden diese Schritte gezeigt und die Möglichkeiten des Studios bei der Analyse vorgeführt. Sicherlich können in der verfügbaren Zeit dabei nicht sämtliche Optionen demonstriert werden, aber der Zuhörer bekommt einen ersten Einblick in die Funktionsweise und die Möglichkeiten von Endeca.

Demonstration durchführen
Im Rahmen der Demo werden, wie im Aufbau beschrieben, Daten aus Twitter mit weiteren Quellen mittels der Integrations Suite (ETL) in den Endeca Server (MDEX) geladen und dort für die Analyse aufbereitet. Durch Hinzufügen von weiteren Informationen kann dann mit dem Studio ein erste Analyse dieser Daten erfolgen und in einem Dashboard angezeigt werden.

Abschließende Diskussionsrunde
Mit Endeca hat Oracle im BI Umfeld nun eine Plattform für hoch dynamische analytische Lösungen. Durch die InMemory Technik und die schnelle Anpassbarkeit des Endeca Information Discovery bieten es dem Anwender neue Wege zur übergreifenden Analyse von strukturierten und unstrukturierten Daten. Er ist mit Endeca eigenständig in der Lage die gewünschten Informationen schnell und einfach miteinander zu verknüpfen. Durch die vielfältigen Funktionalitäten kann der Anwender recht einfach zu neuen Analyseergebnissen gelangen. Hierin liegen die Stärken von Endeca.

Durch diesen Vortrag über Endeca und vor allem mit der kurzen Demo wird ein erster, kurzer Einblick in die Funktionsweise und die Möglichkeiten der EndecaProdukte gegeben. Gerade die schnelle Umsetzung bei der Verknüpfung von Daten, strukturierter wie unstrukturierter, bietet neue Formen der Aufbereitung für Endanwender. Ein Einsatz in Kombination mit Social Media-Daten bietet sich geradezu an. In der abschließenden Diskussionsrunde kann ein erster Austausch mit den Zuhörern untereinander zu möglichen Einsatzszenarien von Endeca erfolgen.

Für die Zukunft ist die Produktpolitik von Oracle für Endeca interessant. Vor allem wichtig wird sein, wie gut die Integration von Endeca in die bestehende Oracle BI-Suite gelingen wird.

Kontaktadresse:
Norbert Henz

Trivadis GmbH
Paul-Dessau-Straße 6
D-22761Hamburg

Telefon: +49 (0) 40-248 591 30
Fax: +49 (0) 40-248 591 59
E-Mail norbert.henz@trivadis.com

Internet: www.trivadis.com
Twitter @Trivadis