Vor einigen Monaten wurde die Übernahme der Firma Endeca durch Oracle bekanntgegeben. Der Produktname leitet sich interessanterweise aus dem deutschen Wort „entdecken" ab. Dieser Artikel gibt einen Überblick über das Produkt „Oracle Endeca Information Discovery" und zeigt Möglichkeiten auf, die sich durch diese Technolgie bieten.

Informationen mit Oracle Endeca Information Discovery entdecken

Mathias Klein, ORACLE Deutschland B.V. & Co. KG

In globalen Unternehmen müssen Fachanwender Tag für Tag wichtige und unternehmenskritische Entscheidungen treffen und benötigen für ihre komplexen Fragestellungen die Transparenz aller relevanten Informationen. Die enorme Menge und Vielfalt an Daten, die heutzutage in unserer Informationsgesellschaft entstehen, stellt Unternehmen und deren IT-Abteilungen vor große Herausforderungen. Häufig sind die wesentlichen Informationen über verschiedene Systeme verteilt und werden für übergreifende Fragestellungen manuell zusammengeführt und zeitaufwändig ausgewertet. Diese Informationen können in den unterschiedlichsten Formaten vorliegen (strukturiert, halbstrukturiert, unstrukturiert) sowie in den verschiedensten Systemen gespeichert sein (Data Warehouse, interne Datenbanken, Office-Dokumente). Zudem entstehen durch die rasante Entwicklung des Internets neue externe Informationsquellen, die für eine Auswertung, vor allem in Kombination mit internen Daten, interessant sein können (Blogs, Facebook, Twitter etc.).

Business Intelligence ist weit verbreitet, wenn klar formulierte Fragestellungen bestehen, der Datenumfang eindeutig definiert ist und dafür ein passendes Datenmodell erstellt wurde. Allerdings bieten diese traditionellen BI-Technologien nicht die notwendige Agilität, um effizient auf die ständig wechselnden Fragestellungen der Fachbereiche zu reagieren sowie die rasche Integration von neuen Datenquellen zu gewährleisten. Weiterhin ist die Verwendung dieser Tools meist nur speziell geschulten Anwendern möglich und bedarf bei neuen Anforderungen der Unterstützung von IT-Spezialisten, um neue Reports und Auswertungen zu erstellen.

Durch die Übernahme von Endeca hat Oracle vor einigen Monaten eine Technologie hinzugekauft, die die beschriebenen Problemstellungen abdecken kann. Endeca wurde ursprünglich mit dem Ziel entwickelt, Anwender im Internet schnell und komfortabel zu den gewünschten Informationen oder Produkten zu führen. Neben umfangreichen Suchfunktionen über Frei-texte ist ein zentrales Element die von Endeca entwickelte geführte Navigation („Guided Navigation"), die in den meisten Online-Shops mittlerweile zur Standardfunktionalität gehören. Endeca war einer der Vorreiter auf diesem Gebiet und ist heute vor allem in Nordamerika Marktführer in diesem Bereich. Aus dieser Technologie heraus entwickelte sich Oracle Endeca Information Discovery (OEID). Es ermöglicht Fachanwendern in Unternehmen, selbstständig und ohne tiefgreifende IT-Kenntnisse an die gewünschten Informationen oder Analyseergebnisse zu gelangen.

OEID kombiniert Funktionalitäten einer Suchmaschine mit der Leistungsfähigkeit von analytischen BI-Tools. Es basiert auf einer facettierten Datenhaltung und ist für verschiedenste Anwendungsfälle in der Industrie, im Handel und bei Behörden im Einsatz. Dieser grundlegend neue Ansatz der Datenhaltung erfordert kein vordefiniertes Datenbankschema, sondern Datensätze werden als Sammlung von Key-Value-Paaren gespeichert. Jeder Datensatz kann anders aufgebaut sein und das Datenmodell wird aus den geladenen Daten abgeleitet. Aufgrund dieser Charakteristik können Anwendungen sehr schnell implementiert und iterativ weiterentwickelt werden.

Ein typischer Anwendungsfall ist beispielsweise die Analyse der Gewährleistungskosten bei einem Automobilhersteller. Daten aus verschiedenen Systemen werden so zu einem Gewährleistungs-Datensatz zusammengefügt, der iterativ erweitert werden kann:

- Gewährleistungs-Informationen: Welcher Befund wurde festgestellt und welche Kosten sind entstanden?
- Fahrzeug-Konfiguration: Mit welcher Konfiguration wurde das Fahrzeug ausgeliefert?
- Händler-Informationen: Wo wurde das Fahrzeug repariert?
- Teile-Informationen: Welche Teile werden häufig ersetzt?
- Lieferanten-Informationen: Welcher Lieferant hat defekte Teile geliefert?
- Bonitäts-Informationen: Existiert ein Zusammenhang zwischen der Bonität eines Lieferanten und der Qualität der gelieferten Teile?
- Informationen aus dem Internet und sozialen Medien: Welche Qualitätsprobleme werden von Kunden in Internetforen diskutiert?

Antworten auf diese Fragestellungen können durch Mitarbeiter einer Fachabteilung selbstständig mithilfe einer Endeca-Anwendung erlangt werden. Analog lassen sich iterativ weitere Datenquellen und -felder zu einem Endeca-Datensatz hinzufügen.

Funktionsweise

Endeca erlaubt eine Vielzahl von Abfragemöglichkeiten wie Navigation, interaktive Visualisierungen, Analysen, Bereichsfilter, Geodatenfilter und darüber hinaus andere Abfragetypen, die in der Regel nicht in traditionellen BI-Tools Verwendung finden, etwa Volltextsuche oder Geo-Analysen wie Umkreissuche und Bereichsfilter in Karten.

Jedes Attribut, das in den Datensätzen enthalten ist, kann als Filter-Kriterium dienen. Dabei funktionieren diese Abfragen gleichermaßen für strukturierte, halbstrukturierte und unstrukturierte Inhalte, die im Endeca-Server gespeichert sind. Ergebnisse von Abfragen können wie bei einer Suchmaschine mit einer Ergebnisliste beantwortet werden, wobei dem User das für ihn interessanteste Ergebnis durch Konfiguration von Relevance-Ranking-Modulen zuerst präsentiert werden kann. Alle Charts und Filtermöglichkeiten in der Anwendungsoberfläche berechnen sich nach jedem Filter neu und die Faceted Navigation zeigt dem User nur die aktuell gültigen Navigationsoptionen an. So werden Ergebnisse immer neu zusammengefasst präsentiert, sodass die Nutzer einen Anhaltspunkt haben, wie sie die Ergebnisse weiter verfeinern und erkunden können. Der Anwender kann die Zusammenfassungen und Filter weiterverwenden, ohne dazu komplexe SQL-Abfragen erstellen zu müssen. Filter können einfach durch Klicken hinzugefügt oder gelöscht werden.

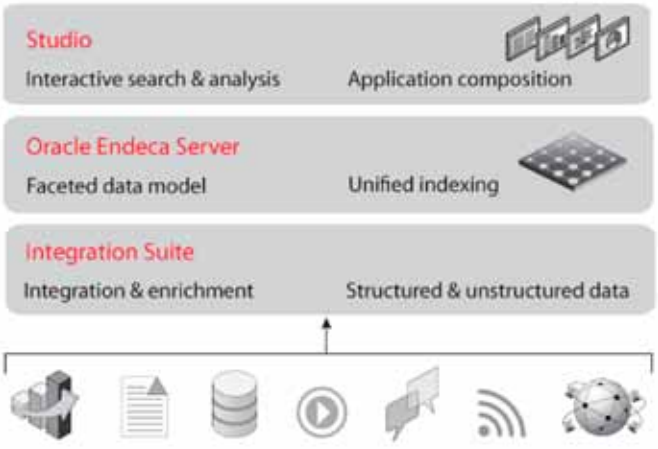

Abbildung 1: Komponenten von Oracle Endeca Information Discovery

Abbildung 2: Beispielhafte Anwendung auf Basis von OEID

Oracle Endeca Server

Die zentrale Komponente in OEID ist der Endeca-Server, eine spaltenorientierte In-Memory-Datenbank, die gleichermaßen Such- und Analysefunktionen unterstützt (siehe Abbildung 1). Diese ähnelt in vielerlei Hinsicht modernen Datenbank-Systemen, wurde jedoch speziell für die Besonderheiten der übergreifenden Analyse von unstrukturierten, halbstrukturierten und strukturierten Daten entwickelt. Im Mittelpunkt steht eine spaltenorientierte Datenhaltung, die hohe Performance und gute Skalierbarkeit ermöglicht. Diese Struktur erlaubt eine starke Komprimierung aufgrund der Gleichartigkeit der Daten innerhalb der Spalten. Dank der geringen Speicherbelastung erfolgt die Ergebnisbereitstellung besonders schnell. Jede Informationsspalte wird sowohl auf dem Datenträger als auch im Arbeitsspeicher gesichert. Die Datensätze werden dabei einmal nach dem Wert und ein zweites Mal nach der universellen Datensatz-ID sortiert. Jede Spalte enthält zudem einen Index mit Baumstruktur, der im Arbeitsspeicher zwischengelagert wird, um die Suche und Bereitstellung der im Endeca-Server enthaltenen Daten zu beschleunigen.

Neben der schnellen Filterung und Exploration bietet der Endeca-Server als weitere zentrale Funktionalität die Möglichkeit, Ad-hoc-Abfragen über eine integrierte Analyse-Sprache zu erstellen. Die Endeca Query Language (EQL) bietet analytische Funktionen in SQL-ähn-

licher Syntax zur flexiblen Aggregation von Informationen, um Trends, Statistiken, analytische Visualisierungen und Vergleiche in Analyse-Anwendungen darzustellen. Sie unterstützt den Umgang mit verschiedenen Datentypen wie numerische, Datums- und Uhrzeitwerte. In Anwendungen können dadurch Zeitdaten verwendet und zeitbasierte Sortier-, Filter- und Analyse-Vorgänge durchgeführt werden. Um eine hohe Auslastung von Multicore-CPU-Systemen zu erreichen, wird die Berechnung einzelner EQL-Queries auf die verschiedenen Prozessoren verteilt und parallel verarbeitet. Die Kommunikation mit dem Endeca-Server erfolgt über Webservices. Sowohl zum Beladen mit neuen Daten als auch für die Abfrage von Informationen stehen standardisierte Schnittstellen zur Verfügung. Zudem existiert für große Datenmengen ein Bulk-Loader-Interface. Während des Betriebs können neue Daten zum Index hinzugefügt oder bereits gespeicherte Informationen aktualisiert werden, ohne dass eine Neuindexierung aller Daten erforderlich ist.

Oracle Endeca Studio

Endeca Studio bietet die Möglichkeit, interaktive Anwendungen auf Basis des Oracle-Servers zu entwickeln. Es basiert auf einer webgestützten Infrastruktur, auf die Endanwender über einen Browser zugreifen können. Verschiedene vorgefertigte Komponenten können per „Drag & Drop“ auf die Oberfläche gezogen und dort konfiguriert werden. So lassen sich in kurzer Zeit neue Anwendungs-Oberflächen entwickeln und einer breiten Anwenderzahl zur Verfügung stellen (siehe Abbildung 2). Es werden folgende Komponenten angeboten:

- Filterkomponenten, um Daten zu durchsuchen
  - Breadcrumbs
  - Guided Navigation
  - Range Filters
  - Search Box
- Visualisierungskomponenten, um eine detailliertere Sicht auf die Daten zu ermöglichen
  - Alerts
  - Chart
  - Compare
  - Cross Tab
  - Map
  - Metrics Bar
  - Tag Cloud
- Die Komponenten zur Ergebnisanzeige
  - Data Explorer
  - Record Details
  - Results List
  - Results Table

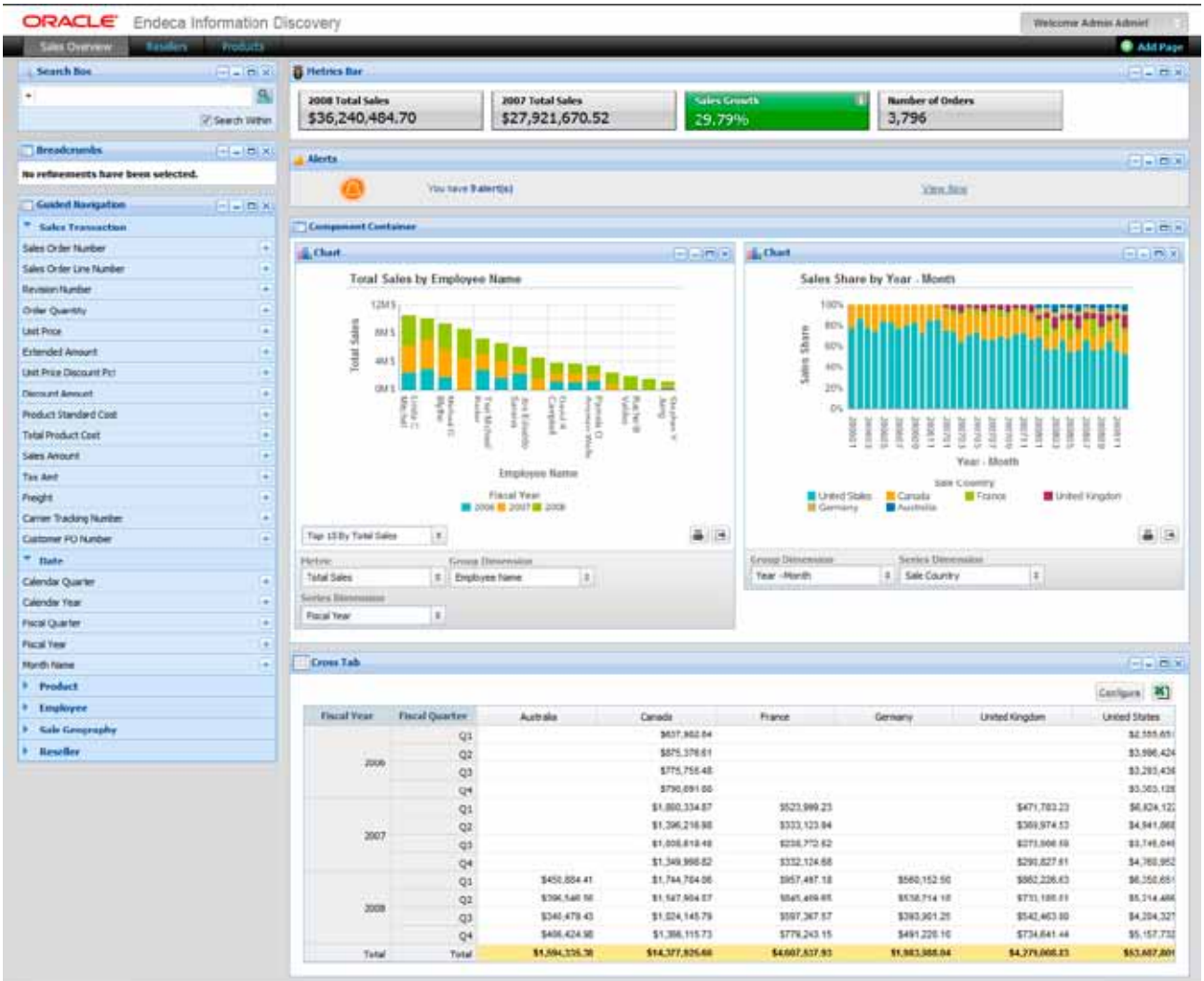

Abbildung 3: Analyse von Social-Media-Daten

Über ein Java-API ist es zudem möglich, weitere Visualisierungs- und Filterkomponenten zu entwickeln.

Oracle Endeca Integration Suite

Um Daten aus verschiedenen Quellsystemen in den Endeca-Server zu laden, besteht die Integration Suite aus einer breiten Palette an leistungsfähigen ETL-Tools, System-Konnektoren und Content-Enrichment-Bibliotheken für die Zusammenführung und Anreicherung vielfältiger Informationen. Sie ermöglicht die effiziente Vernetzung von strukturierten und unstrukturierten Daten zu einer einheitlichen, integrierten Sicht. Die Integration Suite setzt sich im Einzelnen aus den folgenden Komponenten zusammen:

- *Integrator*
  Eine umfassende ETL-Umgebung, die Daten unter anderem aus relationalen Datenbanken, XML- oder Excel-Dateien extrahieren kann. Das Beladen und Updaten des Endeca-Servers kann durch einen Scheduler zeitgesteuert ablaufen.
- *Content Acquisition System*
  Eine Crawling-Umgebung, die verschiedene Konnektoren zur Integration unstrukturierter Daten bietet (etwa Crawling von Office oder von PDF-Dokumenten aus Dateisystemen) sowie Anbindungen an bestehende Content Management Systeme (CMS) ermöglicht. Beim Crawlen von Dokumenten auf Dateisystemen werden Zugriffsberechtigungen mitextrahiert und können in einer Endeca-Anwendung zum Einsatz kommen. Zum Leistungsumfang zählt auch ein Webcrawler zur Anbindung von Internet-Foren, Twitter oder Facebook.
- *Text Enrichment und Sentiment-Analyse*
  Optional können Text-Analyse- und Text-Mining-Produkte eingebunden werden, um wichtige Begriffe (wie Personen-, Orts- und Firmen-Namen) aus textbasierten Informationsquellen zu extrahieren sowie die positive oder negative Tonalität (Sentiment-Analyse) eines Forenbeitrags zu erkennen. Diese zusätzlichen Informationen können in einer Endeca-Anwendung für Analyse-Zwecke herangezogen werden.
  Abbildung 3 zeigt beispielhaft, wie eine solche Social-Media-Applikation aussehen könnte, die anhand von Kunden-Kommentaren erstellt wurde.

Fazit

Oracle Endeca Information Discovery bietet eine umfassende Plattform zur Bereitstellung von analytischen Anwendungen. Es eignet sich vor allem für Anwendungsfälle, bei denen Daten aus den verschiedensten Systemen in unterschiedlichen Formaten vernetzt analysiert werden müssen. Der aufwändige Planungs- und Modellierungsprozess traditioneller Tools entfällt weitestgehend, was eine kurze Implementierungsdauer von wenigen Wochen ermöglicht. OEID versetzt Anwender in die Lage, mit einem einfach zu verwendenden Analysewerkzeug schnell und selbstständig an alle relevanten Informationen zu gelangen. Dies verringert sowohl Aufwände als auch Abhängigkeiten von der IT-Abteilung und hilft, den ständig wachsenden Geschäftsanforderungen der Fachbereiche gerecht zu werden.

Weitere Informationen

1. Oracle Endeca Information Discovery Produktinformationen: http://www.oracle.com/technetwork/middleware/endeca/overview/index.html
2. Oracle Endeca Information Discovery Channel auf YouTube: http://www.youtube.com/user/OracleEID/featured

Mathias Klein
mathias.klein@oracle.com