Unternehmen stützen sich seit Jahrzehnten bei ihren Geschäftsentscheidungen auf Transaktionsdaten, die in relationalen Datenbanken gespeichert sind. Neben diesen kritischen Daten gibt es aber noch eine Vielzahl weiterer Quellen mit zum Teil weniger streng strukturierten Daten wie Office-Dokumenten, E-Mails, Beiträgen aus Internet-Foren, Blogs, sozialen Netzwerken oder Sensordaten. Durch Anbindung dieser meist brachliegenden Datenquellen lassen sich nützliche Zusatz-Informationen gewinnen und für die ganzheitliche Darstellung geschäftlicher Zusammenhänge einsetzen.

Aufbau agiler BI- und Discovery-Applikationen mit Oracle Endeca

Harald Erb, ORACLE Deutschland B.V. & Co. KG

Für den Aufbau agiler BI- und Discovery-Applikationen stellt der Artikel das neue Produkt Oracle Endeca Information Discovery (OEID) vor, setzt es anhand eines durchgängigen Beispiels in den Gesamtkontext von Oracles Business-Analytics-Strategie beziehungsweise des zugehörigen Lösungsangebots und erläutert, wie OEID auf neuartige Weise Funktionen einer Suchmaschine mit der Leistungsfähigkeit eines Business-Intelligence-Werkzeugs kombiniert.

Social Media Monitoring – ein Laborbeispiel

Zu den häufig diskutierten Anwendungsbereichen von Discovery Applikationen gehört das systematische Beobachten und Analysieren von Social-Media-Beiträgen und Dialogen in Diskussionsforen, Weblogs, Communities etc., um unter anderem die Zustimmung oder Ablehnung der Konsumenten zu Produkten und Services besser verstehen zu lernen. Nachfolgend wird auf Basis der Oracle-Business-Analytics-Plattform ein vereinfachtes Szenario beschrieben, das in vier Schritten („Aquire“, „Organize“, „Analyze“ und „Decide“) den Aufbau einer Analyse-Applikation skizziert. Die Oracle-Business-Analytics-Plattform bietet für die Umsetzung solcher Vorhaben die passende Infrastruktur und besteht in unserem Szenario aus folgenden sogenannten „Engineered Systems“ (siehe Abbildung 1):

- *Big Data Appliance*
  Zur Bereitstellung der zu verarbeitenden Massendaten (auch in unbeziehungsweise semistrukturierter Form)
- *Exadata*
  Für die kombinierte Analyse von Big Data mit den traditionellen Unternehmens-Datenquellen wie Data Warehouse, OLTP-Datenbank.
- *Exalytics*
  Für den Aufbau/Betrieb analytischer In-Memory-Applikationen

In unserem Laborbeispiel gehen wir von bereits vorhandenen Unternehmens-Datenquellen (Data Warehouse, PLM-System etc.) aus, Twitter-Kurznachrichten sollen als zusätzlich anzubindende Datenquelle die Basis für Social-Media-Analysen bilden, die später mit OEID erfolgen.

Aquire

In unserem Beispiel müssen zunächst Twitter-Daten – eingeschränkt nach eigenen Suchbegriffen – beschafft und

Abbildung 1: Die Oracle-Business-Analytics-Plattform

Performance by Design.

areto kennt die Stellschrauben.

ORACLE Gold Partner

Wer voraus denkt, ist schneller am Ziel: Mit sorgfältiger Datenmodellierung, versiertem Technologie-Einsatz und nachhaltigen Projektstandards bringen wir Struktur in Ihre Datenbestände. So entstehen genau auf Ihre Anforderungen zugeschnittene BI- und Reporting-Anwendungen mit hoher Akzeptanz und Leistungsfähigkeit.

Entdecken Sie jetzt eine neue Dimension in Business Intelligence und Reporting.

Rufen Sie uns an:
0221 66 95 75-75

areto consulting gmbh · Data Warehouse · Business Intelligence
Julius-Bau-Straße 2 · 51063 Köln · 0221 66 95 75-0 · www.areto-consulting.de

areto
CONSULTING. IT WORKS.

zur Oracle Big Data Appliance (BDA) übertragen werden. Twitter stellt dazu verschiedene Web-Service-APIs bereit (siehe http://dev.twitter.com), die es uns entweder erlauben, per Bulk-Collect (REST-API) historische Tweets über einen Zeitraum von sieben bis zehn Tagen zu erhalten oder durch Nutzung der Streaming-APIs kontinuierlich Tweets zu vorgegebenen Suchbegriffen oder einzelnen Usern in die Big Data Appliance zu laden. Twitter stellt die angeforderten Daten im XML-/JSON-Dateiformat bereit, daher bietet sich für unser Szenario die dateiorientierte Speicherung der Twitter-Inhalte im Hadoop-Distributed-Filesystem (HDFS) der BDA an. Alternativ steht in der BDA für die satzorientierte Speicherung der Social-Media-Daten die Oracle NoSQL Datenbank als Universal-Key-Value-Speicher zur Verfügung, die technisch auf der Oracle Berkeley Datenbank basiert.

Abbildung 2 zeigt, wie die Beschaffung der Twitter-Daten zum Beispiel mit dem Java-Programm „twitter_search.jar" programmatisch umgesetzt werden kann. Möchte man nahezu in Echtzeit User-Statusmeldungen oder die Ergebnisse eigener Suchanfragen aus dem globalen Twitter-Stream abrufen, dann startet man mit dem Aufruf „Stream" einen Twitter-Streaming-Job, der die resultierenden Tweets in ein XML-Dateiformat wandelt und anschließend für die Weiterverarbeitung in einem Eingangsverzeichnis im verteilten Dateisystem (HDFS) der BDA speichert.

Über diesen Weg erhält man zusammen mit der Twitter-Nachricht den Namen des Users, den Zeitstempel der Nachricht sowie einige User-Metriken wie „Anzahl Follower" und „Anzahl Freunde". Alternativ lassen sich per Aufruf „Search" ältere Twitter-Daten mit eigenen Schlagworten durchsuchen. Im Ergebnis erhält man zusätzlich zu den gefundenen Tweets nur die Zeitangabe und den User-Namen des Verfassers. Die User-Metriken (social importance metrics), die wichtig für die Bestimmung des Einflusses des Users in der Netzwelt sind, fehlen allerdings. Über einen User Lookup lässt sich dieses Informationsdefizit jedoch beheben, indem nachträglich die Ergebnisse der Twitter-Suche um diese Metriken ergänzt werden.

Für unser Beispielszenario werden rund um das Thema Mobilfunk die wichtigsten Suchbegriffe in einer einfachen Konfigurationsdatei hinterlegt und kategorisiert. So bilden die Begriffe „@iPhone", „@iPhone3", „@iPhone4" eine Kategorie „IP". Die vom Twitter-Streaming-Job abgerufenen Ergebnisse werden dann bei der Umformatierung in ein XML-Ausgabeformat noch zusätzlich um eine entsprechende Kategorie-Information erweitert. Dieser Verarbeitungsschritt ermöglicht später das Organisieren der Daten mit dem Hadoop-Framework „MapReduce".

Organize

Nach der gezielten Beschaffung der Twitter-Daten findet in unserem Beispiel nun das Framework „MapReduce" seine Anwendung. Ziel ist dabei die Durchführung einer Sentiment-Analyse, um aus den Twitter-Nachrichten systematisch die Konsumenten-Zustimmung beziehungsweise -Ablehnung zu Produkten oder Services ermitteln zu können (siehe Abbildung 3). Die Sentiment-Analyse selbst ist simpel gehalten: Der Text eines Tweets wird tokenisiert und ein Fuzzy-Match-Algorithmus durchsucht dafür hinterlegte Wörterbücher nach passenden positiven oder negativen Wörtern; die Grammatik der Sätze bleibt unberücksichtigt. Wird ein positives Wort gefunden, erhöht sich der Sentiment Score, bei negativen Treffern reduziert er sich entsprechend. Für die Anwendung anspruchsvollerer Methoden bietet BDA zusätzliche In-Database-Funktionen wie Text Mining, Data Mining oder die Statistikumge-

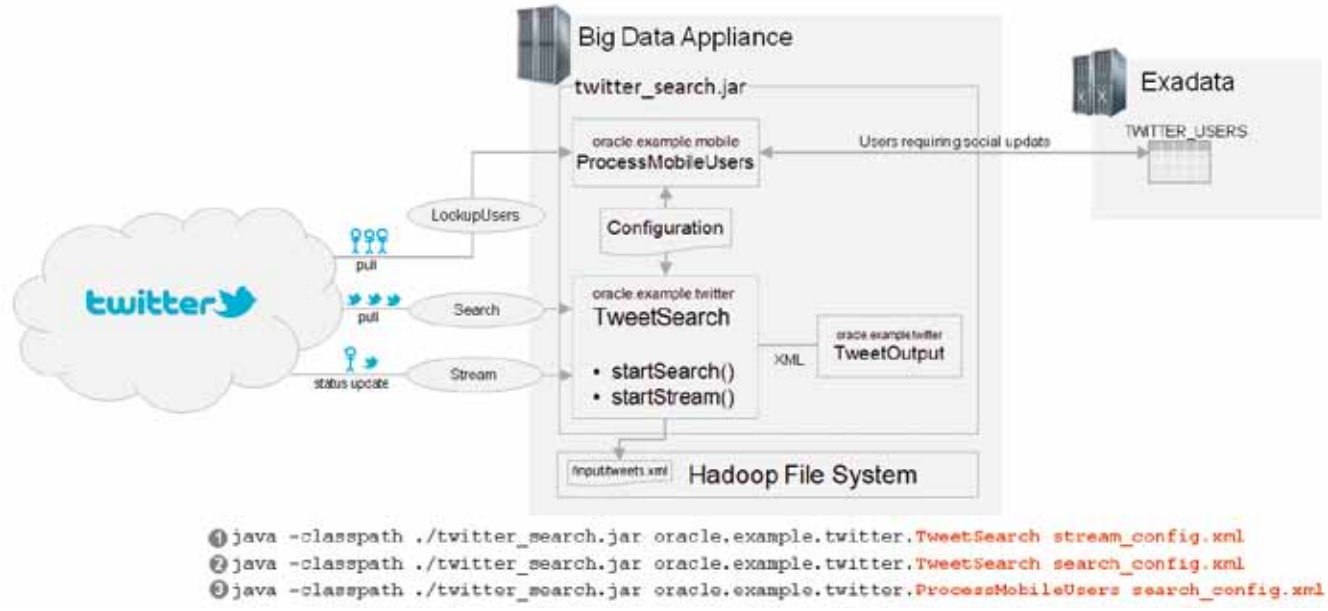

Abbildung 2: Akquisition von Twitter-Daten

Tweet	Sentiment
Would like to take the opportunity to thank @Vodafone customer service for delivering fast at no cost! #yay #awesome.	5
I hate the display. It is broken. Not Good!.	-3

Abbildung 3: Sentiment Score der Tweet-Nachrichten berechnen

bung „Project R“ und unterstützt auch Semantik-Analysen.

Die MapReduce-Jobs fassen in unserem Beispiel in der Mapper-Phase die Tweets und ihre errechneten Sentiment-Scores nach bestimmten Kriterien zusammen, zum Beispiel auf Tagesbasis, nach Telekommunikations-Providern (VF = Vodafone), nach den im Aquire-Schritt festgelegten Tweet-Kategorien (Staff, Contracts, iPhone etc.) oder nach Twitter-Usern. Als Ergebnis produziert der Mapper Key-Value-Paare, die im Anschluss die „Shuffle & Sort“-Phase durchlaufen, bevor sie an den Reducer übergeben werden. In der Abbildung 4 sind die beteiligten Komponenten zu sehen, repräsentiert durch das Java-Programm „mobile_mr.jar“, für unser Beispiel sowie für die in der Mapper-Phase zu erledigenden Aufgaben.

In der Reducer-Phase erfolgt schließlich die Weiterverarbeitung der sortierten Schlüssel und Werte-Arrays. In unserem Beispiel entstehen dabei neue Key-Value-Paare mit nun zwei Metriken: dem aufsummierten Sentiment-Score und der Anzahl der Vorkommnisse pro Schlüssel (siehe Abbildung 5). Der letzte Schritt ist wiederum die Ausgabe der vom Reducer generierten, finalen Key-Value-Paare, die nun als Textdateien im HDFS abgelegt werden.

In unserem Beispiel soll später einmal der Marketing-Bereich eines Telco-Anbieters mit Endeca Information Discovery die Tonalität der Twitter-Posts seiner Konsumenten für eine gezielte Kunden-Ansprache nutzen können. Dafür werden die Ergebnisse der MapReduce-Jobs mit dem Oracle Direct Connector for HDFS (ODCH) via InfiniBand-Netzwerkverbindung in eine Oracle-Datenbank geladen (siehe Abbildung 6).

Die im verteilten Dateisystem der BDA abgelegten Key-Value-Paare können somit in einem Exadata Data Warehouse über eine externe Tabelle bequem mit SQL abgefragt und weiterverarbeitet werden. Für unser Beispiel ist es ein wichtiger Punkt, wenn es gelingt, die neuen Erkenntnisse aus dem Twitter-Kanal mit den Kundendaten etwa im Data Warehouse verknüpfen zu können. Ab jetzt können bereits Unternehmensdaten zusammen mit den aufbereiteten Social-Media-Daten per In-Database-Analytik von einer großen Anzahl von Anwendern genutzt werden.

Analyze

In unserem Beispiel sind nun die Sentiments zu allen Tweets errechnet, die

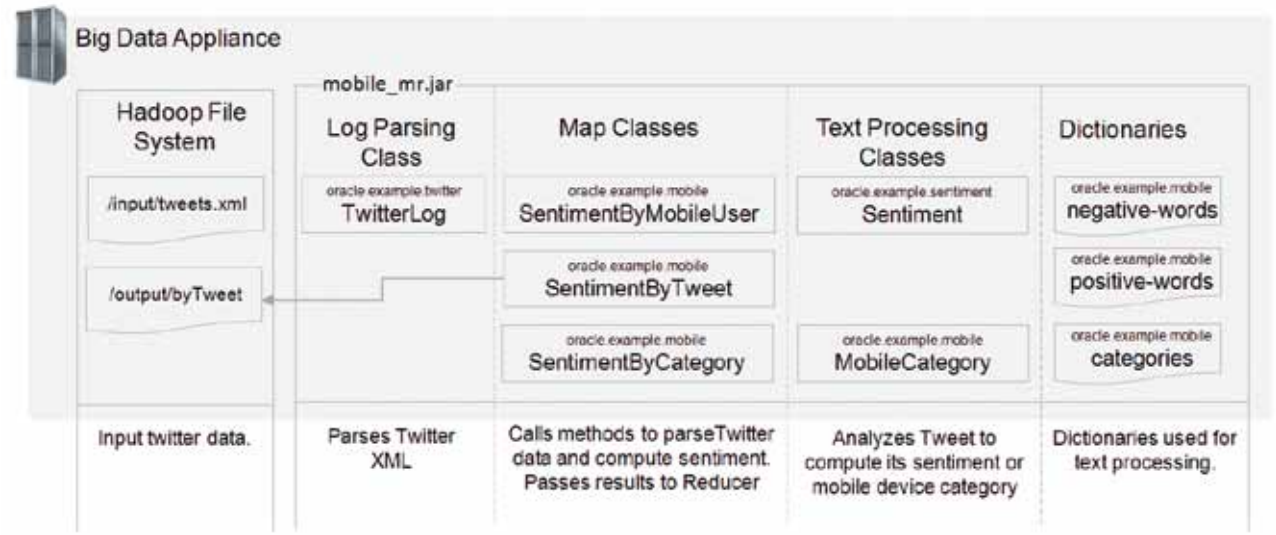

Abbildung 4: Die Mapper-Phase

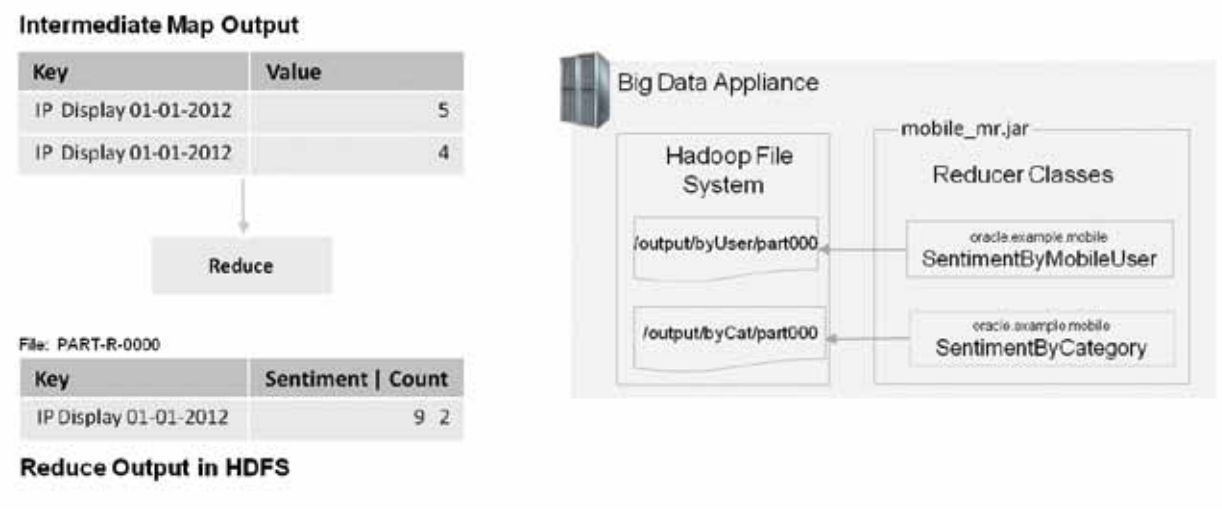

Abbildung 5: Die Reduce-Phase

wir zu diesem Zweck vom globalen Twitter-Stream in die Big Data Appliance übertragen und anschließend nach verschiedenen Kategorien („auf Tagesbasis", „nach Hersteller", „nach Mobilfunkkunden", „nach Service-Aspekten") aggregiert haben. Ferner kennen wir aufgrund einiger Metriken („Anzahl Freunde", „Anzahl Follower") die „Social Importance" der Verfasser zu den untersuchten Tweets.

Durch das Laden dieser Twitter-Daten in das Enterprise Data Warehouse (Exadata) und die Verknüpfung mit den vorhandenen Unternehmensdaten lässt sich nun beispielsweise die Kundenzufriedenheit über die Zeit darstellen. Es können aber auch Trends ermittelt und die Zufriedenheit der eigenen Kunden mit denen des Wettbewerbs verglichen werden. Mit den bekannten analytischen Fähigkeiten der Oracle-Datenbank lassen sich so folgende Fragen beantworten: „Welche ökonomisch wichtigen Kunden sind aufgrund bestimmter Missstände zunehmend frustriert?", „Bei welchen Kundenverträgen ist mit höherer Wahrscheinlichkeit mit einer Kündigung zu rechnen?" oder „Welche Kunden sind – im Social-Media-Kontext betrachtet – Meinungsmacher und nehmen Einfluss auf mein Geschäft?"

Zur Umsetzung der meisten Fragestellungen würde man den klassischen Data-Warehouse- und Business-Intelligence-Ansatz wählen und für die Fachanwender – vom zentralen Enterprise Data Warehouse (Datenmodell) ausgehend – themenspezifische Data Marts (relational/multidimensional) ableiten, die für einen definierten Zeitbereich verdichtete Informationen enthalten. Auf diesen Auswerte-optimierten Datenquellen lässt sich dann In-Memory mit der Exalytics BI Machine auf Basis der Business Intelligence Foundation Suite eine unternehmensweit einsetzbare BI-Plattform aufbauen, die es den Fachanwendern erlaubt, per Self-Service-BI alle relevanten Geschäfts-Informationen abzurufen oder grafisch unterstützt eigene Analysen durchzuführen. Hier stellen sich nun die Fragen: „Wozu also jetzt Endeca einsetzen?" und „Wozu dieser Artikel über agile BI- und Discovery-Applikationen?"

Aus Oracle-Sicht kann man zwischen zwei Arten von Fragestellungen unterscheiden, mit denen analytische Applikationen umgehen müssen. Zum einen gibt es den Typ von Business-Fragestellungen, bei dem im Voraus die entsprechenden Geschäftsprozesse und die dazu benötigten Daten durch die Fachseite bekannt sind: „Wie stellt sich die Umsatzprognose nach Region für einen bestimmten Zeitraum dar?" oder „Wie ist die Performance meiner Organisation im Vergleich zu den gesetzten Zielen?" Zum anderen gibt es Fragestellungen, bei denen weder der entsprechende Geschäftsprozess noch die benötigten Daten vorab durch die Fachseite definiert werden können: „Auf welche Kunden sollen wir uns fokussieren?" oder „Warum gehen meine Verkaufszahlen zurück?" Interessant ist dabei zu sehen, dass der zweite Fragentyp aufgrund seines offenen Charakters im Vergleich zum ersten Typ viel kurzlebiger ist und eher neue Fragen hervorbringt, als abschließend beantwortet zu werden.

Das Interaktionsmodell für die bekannten Fragestellungen kann man ganz gut mit dem Betrachten von aufbereiteten Informationen in einem Standardbericht oder einem BI-Dashboard beschreiben – so wie es heute mit traditionellen Business-Intelligence-Mitteln umgesetzt wird. Bei den heute noch unbekannten, aber morgen schon von den Fachanwendern nachgefragten Analysen ist dagegen ein Interaktionsmodell erforderlich, das eher die Datenerkundung beziehungsweise das Entdecken neuer Zusammenhänge (Data Discovery) unterstützt. Betrachtet man dort zusätzlich den Aspekt der Datenmodellierung, so finden wir bei Business-Intelligence-Lösungen in der Regel den allumfassenden semantischen Layer vor, dessen Aufbau und Pflege Zeit und Geld kostet.

Investitionen dieser Art werden von Unternehmen nur getätigt, wenn sich die Anstrengungen durch Effizienzgewinne bei der Informationsbeschaffung wieder amortisieren. Gleichzeitig sinken weiterhin die Kosten für Speichermedien und mit der Popularität von Hadoop steigen die Aussichten,

dass aus nichtmodellierten Daten Nutzen gezogen werden kann.

Aus diesen beiden Blickwinkeln erkennt man, dass sich traditionelle Business-Intelligence- und Data-Discovery-Lösungen ergänzen können. Die nach den Anforderungen der Fachseite aufgebaute Business-Intelligence-Anwendung liefert qualitätsgesicherte Ergebnisse für bekannte Fragestellungen. Es können aber auch neue Fragen aufgeworfen werden, deren Beantwortung erst mit einem neuen Release der BI-Anwendung oder gar des Data Warehouse möglich ist – im schlimmsten Fall ist die Anforderung bis dahin schon obsolet geworden. Mit Endeca sind dagegen neue fachliche Fragestellungen schneller zu beantworten, insbesondere dann, wenn die dafür notwendigen Informationen in den unterschiedlichsten Formaten vorliegen (strukturiert, semistrukturiert, unstrukturiert) und in den verschiedensten Systemen gespeichert sind (DWH, operative Datenbanken, Office-Dokumente). Stellt sich bei der Arbeit mit Endeca Information Discovery heraus, dass die beantworteten Fragestellungen regelmäßig benötigt werden, kann dies in die Release-Planung für die nächste DWH-Version beziehungsweise die Version der Business-Intelligence-Applikation einfließen.

Zurück zum Beispiel: Abbildung 7 zeigt den typischen Endeca-Fall. Ein Großteil der zu analysierenden Daten (einschließlich der aufbereiteten Tweets) stammen aus dem Data Warehouse (Exadata), weitere Zusatz-Informationen liefern in strukturierter Form die Datenbank eines Product-Lifecycle-Management-Systems (detaillierte Gerätebeschreibungen) und in semistrukturierter Form (Vertragsdokumente) ein Content-Management- oder CRM-System. In einer Integrationsphase werden Daten aus unterschiedlichen Quellen miteinander verknüpft und in der Exalytics-Umgebung als denormalisierter „Endeca Record“ im Endeca-Server, einer spaltenorientierten In-Memory-Datenbank, gespeichert.

Die facettierte Datenhaltung im Endeca-Server kommt ohne Tabellen und vordefiniertes Datenmodell (Schema)

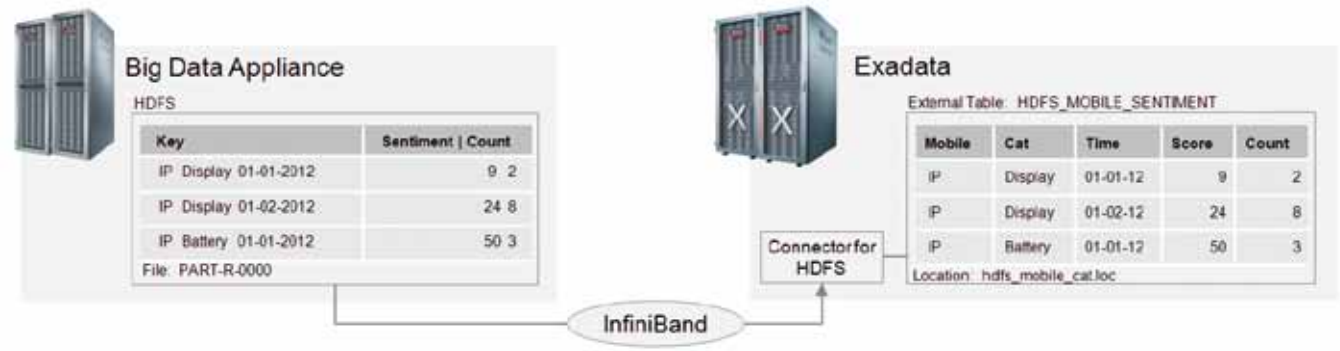

Abbildung 6: Verwendung der MapReduce-Ergebnisse in der Datenbank

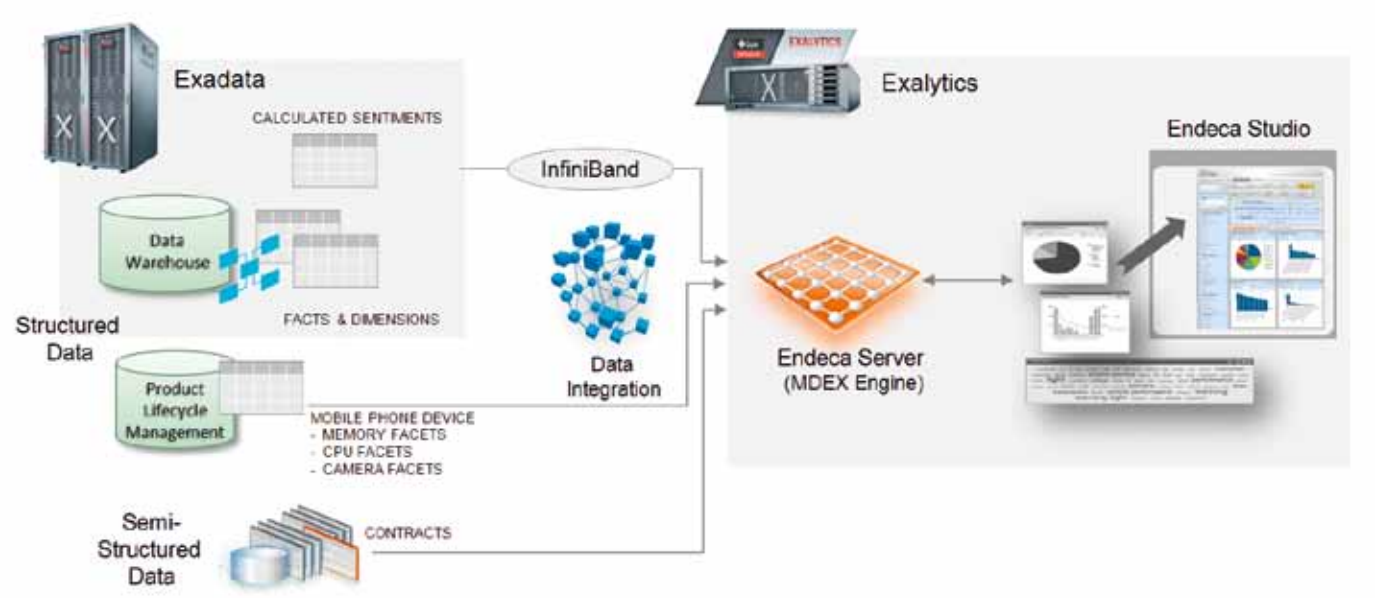

Abbildung 7: „Model as you go“-Ansatz mit Endeca Information Discovery

aus. Die Datensätze selbst werden als Sammlung von Key-Value-Paaren gespeichert, dabei kann jeder Datensatz anders aufgebaut sein – das Facetten-Datenmodell leitet sich automatisch aus den geladenen Daten ab. Das bedeutet zum Beispiel, dass es Daten-Attribute gibt, die exklusiv nur im Data Warehouse, im Product Lifecycle Management System (PLM) oder in den (Meta-)Daten der geladenen Dokumente vorkommen. Andere (globale) Attribute finden sich in mehreren oder allen Datenquellen wieder. Ferner lässt der Endeca-Server auch die Speicherung von semi-strukturierten Daten und Multi-Value-Feldern zu.

Analytische Anwendungen können auf diese Weise schnell implementiert und iterativ weiterentwickelt werden. So kann sich die Erweiterung der Produkt-Dimension in einem relationalen Data-Warehouse-Schema beim ständigen Hinzufügen neuer Produkt-Facetten auf die Dauer als komplex erweisen. Der Endeca Data Store lässt sich dagegen aufgrund seiner flachen, XML-ähnlichen Struktur, mit sich selbst beschreibenden Key-Value-Paaren, beliebig erweitern. Zum Laden verschiedener Datenquellen kommt die Endeca Integration Suite zum Einsatz, die aus dem Werkzeug „CloverETL“ mit Konnektoren und den Content-Enrichment-Bibliotheken für die Zusammenführung und Anreicherung vielfältiger Informationen besteht.

Die Endeca Integration Suite ermöglicht die effiziente Vernetzung strukturierter und unstrukturierter Daten zu einer einheitlichen, integrierten Sicht. Die Kommunikation mit dem Endeca-Server erfolgt über Web-Services, für große Datenmengen gibt es ein Bulk-Loader-Interface. Während des Betriebs können neue Daten zum Endeca Data Store hinzugefügt oder vorhandene Datensätze aktualisiert werden, ohne dass eine Neu-Indexierung aller Daten erforderlich ist.

Für unser Beispiel ist das zur Integration Suite zugehörige „Content Acquisition System“ (CAS) interessant. Dabei handelt es sich um eine Crawling-Umgebung, die verschiedene Konnektoren zur Integration unstrukturierter Daten bietet – in unserem Fall zum Erfassen der Vertragsdokumente im MS-Office- oder PDF-Format. Zum weiteren Leistungsumfang zählt auch ein Webcrawler zur Anbindung von Internet-Sites.

Die Endeca Integration Suite ermöglicht optional auch die Einbindung von Text-Analyse- und Text-Mining-Produkten von Drittanbietern. Auf diesem Weg lassen sich wichtige Begriffe (wie Personen-, Orts- und Firmen-Namen) aus textbasierten Informationsquellen extrahieren oder Sentiment-Analysen durchführen, um die positive/negative Tonalität eines Forenbeitrags oder die Produktzustimmung / -ablehnung von Konsumenten erkennen zu kön nen.

Für das Beispiel wäre dies also eine Alternative zu unserer selbstprogrammierten Sentiment-Analyse in der Big Data Appliance.

Decide

Wie schon gesagt, kann auf der Exalytics BI Machine neben der BI Foundation Suite auch die gesamte Endeca-Infrastruktur betrieben werden. Dazu gehört als Middleware-Komponente „Oracle Endeca Studio“, eine webbasierte Infrastruktur, auf die Anwender per Browser zugreifen können. Endeca Studio stellt eine Bibliothek mit vorgefertigten Portlets zur Verfügung, die per „Drag & Drop“ auf die Anwenderoberfläche gezogen und dort konfiguriert werden können. Im Ergebnis steht den Endbenutzern eine agile Discovery-Anwendung zur Verfügung, in der jedes Attribut, das in dem Endeca-Datenbestand enthalten ist, als Filter-Kriterium dienen kann. Alle Charts und Filtermöglichkeiten berechnen sich direkt nach jedem weiteren gesetzen Abfragefilter neu, per „Faceted Navigation“ sieht man als Analyst stets die aktuell verfügbaren Navigationsoptionen. So werden Resultate immer neu zusammengefasst präsentiert, die Nutzer bekommen durch die integrierte Volltextsuche in den semi- beziehungsweise unstruktierten Daten neue Anhaltspunkte, wie sie die Ergebnisse weiter verfeinern und neue Zusammenhänge in den Daten erkennen können.

Harald Erb
harald.erb@oracle.com

Weiterführende Informationen

[1] Jean-Pierre Dijcks: Oracle: Big Data for the Enterprise, Oracle White Paper, Januar 2012, http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/1453236.pdf

[2] V. Murthy, M. Goel, A. Lee, D. Granholm, S. Cheung: Oracle Exalytics In-Memory Machine: A brief introduction, An Oracle White Paper, Oktober 2011, http://www.oracle.com/us/solutions/ent-performance-bi/business-intelligence/exalytics-bi-machine/overview/exalytics-introduction-1372418.pdf

[3] o.V.: A Technical Overview of Oracle Endeca Information Discovery, Oracle White Paper, Mai 2012, http://www.oracle.com/us/solutions/ent-performance-bi/oeid-tech-overview-1674380.pdf

[4] M. Klein: Informationen mit Oracle Endeca Information Discovery entdecken, DOAG News 4-2012

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