Integration verteilter Web-Anwendungen und Datenquellen mit dem Endeca Web Acquisition Toolkit

Harald Erb
ORACLE Deutschland B.V. & Co. KG, Frankfurt/Main

Schlüsselworte

Business Analytics, Endeca Web Acquisition Toolkit, Data Reservoir, Synthetic API, Data Warehouse, Data Mashup, Endeca Information Discovery

Einleitung

Das einzig Beständige ist der Wandel: Kritische Informationen, die Unternehmen täglich als Entscheidungsgrundlage benötigen, unterliegen der permanenten Veränderung und sind noch dazu über viele interne und externe Quellen verteilt. Sei es in Dokumenten, E-Mails, auf Portalen und Websites, etc. – überall finden sich relevante Daten, die wertvolle Erkenntnisse für fundierte Geschäftsentscheidungen liefern können.

Technisch betrachtet müssen die zum Teil sehr schwer zugänglichen Informationen zunächst einmal von den verteilten Anwendungen und Datenquellen beschafft werden bevor die eigentliche Weiterverarbeitung im Data Warehouse stattfindet. Als graphisches Entwicklungswerkzeug setzt das Endeca Web Acquisition Toolkit (Endeca WAT) genau an diesem Punkt an, indem es das Erstellen synthetischer Schnittstellen ermöglicht. Z.B. sollen von einer kommerziellen Website Preisdaten und/oder Kundenbewertungen akquiriert werden, für die der Website-Betreiber keine API bereitstellt. Der nachfolgende Artikel bzw. Vortrag skizziert, wie das Endeca Web Acquisition Toolkit Integrationsaufgaben zur Anbindung externer Datenquellen im Rahmen der aktuellen Oracle Information Management Reference Architecture übernehmen kann [1] [2] [3].

Neue externe Datenquellen erschließen und in die kommerzielle Nutzung überführen

Unternehmen, die weiterhin ihre Marktführerschaft behaupten wollen, müssen ständig Innovationen vorantreiben, um neue Geschäftsmöglichkeiten zu erschließen, den Wert bestehender Kunden zu erhöhen, bestehende Geschäftsprozesse weiter zu optimieren, Kosten zu senken, oder neue Märkte zu finden. Wenn Daten das Herz eines faktengestützten Management-Ansatzes sind, dann sollte beim Aufbau einer analytischen Plattform nicht nur auf das Entdecken neuer Zusammenhänge, sondern auch auf die schnelle und effiziente Operationalisierung der Erkenntnisse Wert gelegt werden.

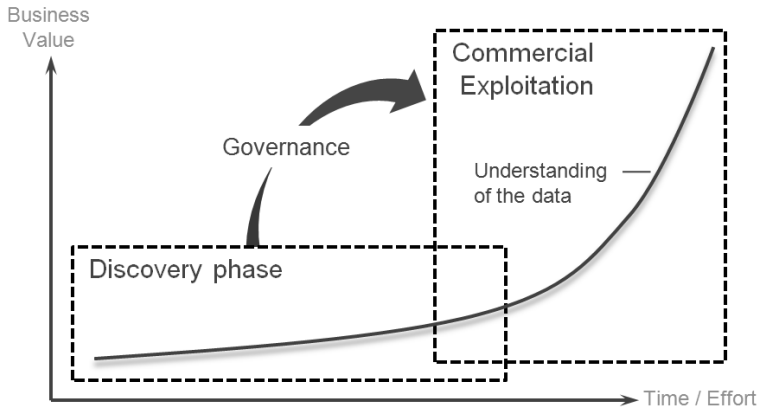

Abb. 1: Operationalisierung der Ergebnisse aus der Discovery Phase

Durch die Verwendung von Big Data Technologien können Unternehmen Daten von geringerer Granularität über eine längere Zeit aufbewahren oder analytische Instrumente - wie zum Beispiel Oracle Endeca - zur Anwendung bringen, deren Einsatz nicht mehr unbedingt das Vorhandensein eines traditionellen relationalen Datenbankmanagementsystems voraussetzt. Ein Data Scientist würde aufgrund bestehender unternehmerischer Herausforderungen und bei Verfügbarkeit neuer Datensets versuchen, diese in einer Self-Service Sandbox mit den internen Unternehmensdaten zu verschneiden bzw. zu kombinieren und dabei eine Reihe unterschiedlicher Techniken und Werkzeuge anwenden. Verwertbare Ergebnisse aus der sog. „Discovery Phase" müssen gemäß Abb. 1 in die zu optimierenden Geschäftsprozesse überführt und auf ihre Alltagstauglichkeit hin überpüft werden. Ziel dabei: Bestätigung der unter Laborbedingungen beobachteten Effekte durch die Business Analysten des Unternehmens, die für ihre Arbeit allerdings eigene Methoden und Werkzeuge zum Einsatz kommen lassen.

Oracle's Information Management Architektur ist in nachstehender Abb. 2 als Konzeptansicht dargestellt und zeigt die Kernkomponenten bzw. wichtigsten Datenflüsse in einer kompakten Form. Hervorgehoben ist dabei, wie Innovationsergebnisse ausgehend von der Datenerkundung im „Discovery Lab" kontrolliert in das Tagesgeschäft überführt werden.

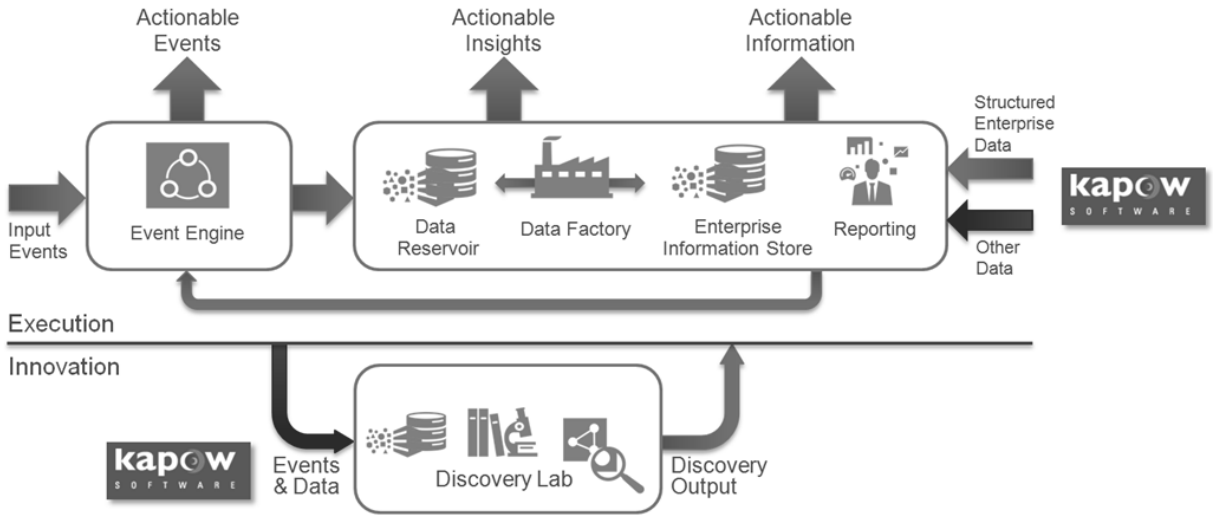

Abb. 2: Endeca Web Acquisition Toolkit („Kapow Softw.") im Kontext der Oracle Information Management Architektur

In dem kompakten, flow-basierten Konzeptmodell sind eine Reihe von Komponenten enthalten, die in vielen Unternehmen zumindest teilweise schon in deren Informationsarchitektur enthalten sind:

Discovery Lab: umfasst abseits der täglichen Datenverarbeitung eigene Datenbestände, Processing Engines und Analysewerkzeuge, die bei der Gewinnung neuer Erkenntnisse Hilfestellung leisten. Das Endeca Web Acquisition Toolkit (basierend auf der OEM Version von Kapow Katalyst) spielt hier als agile Integrationskomponente für externe Datenquellen eine wesentliche Rolle für das Discovery Lab, da es eine schnelle Anbindung solcher Quellen erlaubt.

Data Reservoir: ökonomisch einsetzbares Scale-out Speichersystem, dass massiv parallel Daten verarbeiten kann, ohne dabei allzu strenge Anforderungen an Datenmodellierung oder Formalisierung der Daten zu stellen – typischerweise manifestiert als Hadoop Cluster oder Staging Area in einer relationalen Datenbank.

Enterprise Information Store: groß angelegter Datenspeicher, der unternehmenskritische Daten in formalisierter und modellierter Form beherbergt – typischerweise manifestiert als (Enterprise) Data Warehouse und in Kombination mit dem Data Reservoir als Big Data Management System ausgeprägt.

Reporting: BI Werkzeuge/Infrastrukturkomponenten für zeitgerechtes und akurates Reeporting.

Event Engine: Komponente, die direkt im Datenstrom vordefinierte Ereignismuster erkennt und z.B. per Echtzeitanalyse die nächste beste Aktion/Handlung ermittlet und in einem dauerhaften Speichermedium persistiert.

Discovery Lab Szenario: Kundenrezensionen von Amazun.com akquirieren

Das in der nachstehenden Abb. 3 dargestellte Szenario zeigt, wie in sechs Schritten sämtliche Kundenrezensionen (1) zu einem bestimmten Produkt (hier: Schnappschuss bzw. „Point and Shoot" Kameras) von der E-Commerce Website Amazon.com extrahiert und als (2) Rohdaten in einer Datenbanktabelle zwischengespeichert werden. In einem weiteren Integrationsschritt (3) werden aus den erfassten Fließtexten automatisch per Entitätenextraktion die wichtigsten Themen und per Sentimentanalyse die Tonalität der Beiträge als strukturierte Information im Schema einer Zieldatenbank (4) abgelegt. Die Kombination der neu gewonnenen Amazon-Kundenrezensionen (5) mit den Kamera-Verkaufsdaten aus dem Unternehmensinternen Data Warehouse erfolgt ohne weiteren Modellierungsaufwände per Data Mashup direkt in der Analyse-Anwendung (6).

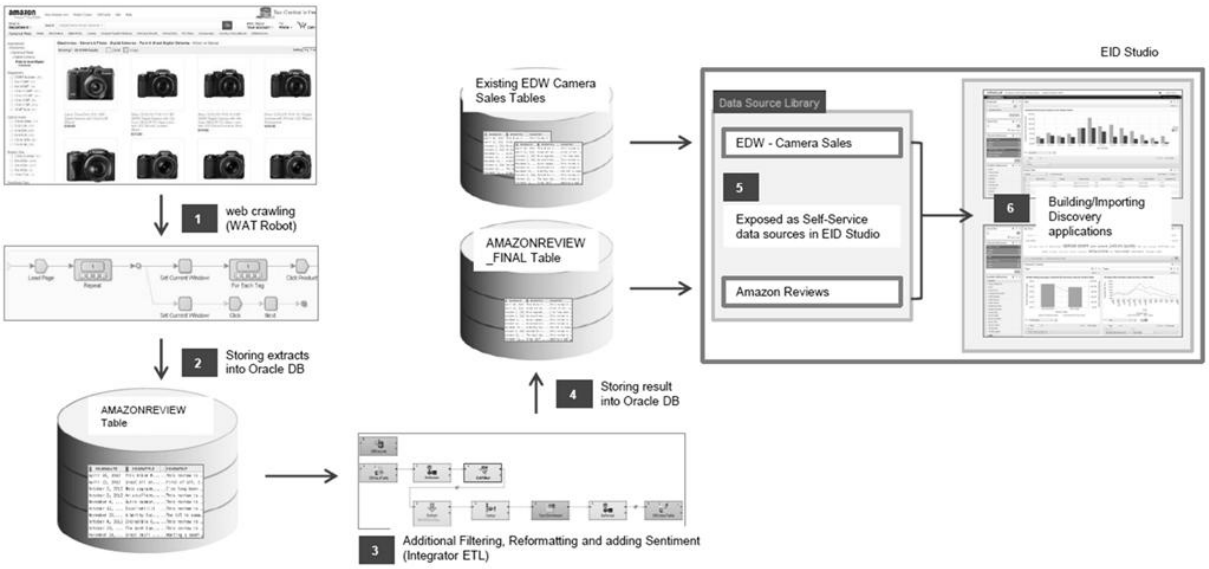

Abb. 3: Verkaufsdaten im Enterpr. DWH zusätzlich mit Kundenrezensionen anreichern

Endeca Web Acquisition Toolkit aka Kapow Katalyst

Kapow Software bietet mit Katalyst eine agile und praxisnahe Lösung zur Integration strukturierter und unstrukturierter Datenquellen: Datenbanken, Standard-APIs, Dokumenten- und E-Mail-Systeme, Web Applikationen, interne sowie Cloud-basierte Anwendungen und Content Management Systeme.

Katalyst besteht aus den drei Hauptelementen [4]

Design Studio, in dem die jeweiligen Integrationsabläufe (aka Robots) erstellt werden,

der **Management Konsole**, in der die Integrationsabläufe angewandt, verwaltet und terminiert und in der die Umgebung überwacht werden kann, sowie

den **RoboServern**, einer cluster-fähigen 24/7-Umgebung, in der die Integrationsbläufe ausgeführt werden.

Das beispiehaft in Abb. 4 dargestellte Kapow Katalyst Design Studio ist eine visuelle, voll integrierte Entwicklungsumgebung. Es verbindet das „Visual Paradigm“ eines Web Browsers in Echtzeit mit einem intuitiven „Visual Flow Editor“. Bei der Erstellung eines Integrationsablaufs navigiert der Anwender über die Quelle (im Szenario Amazon.com), so als würde er ganz normal durch eine Webseite oder ein Excel-Sheet klicken oder einen XML/JSON Editor nutzen. Katalyst Design Studio wurde so ausgelegt, dass nun auch modernste Web Technologien für eine breite Anwendung von Webseiten und Cloud Applikationen unterstützt werden. Da das Design Studio mit Live-Daten arbeitet, sieht der Entwickler sofort, ob die ausgewählten Daten wunschgemäß formatiert und alle erforderlichen Geschäftsregeln eingehalten werden.

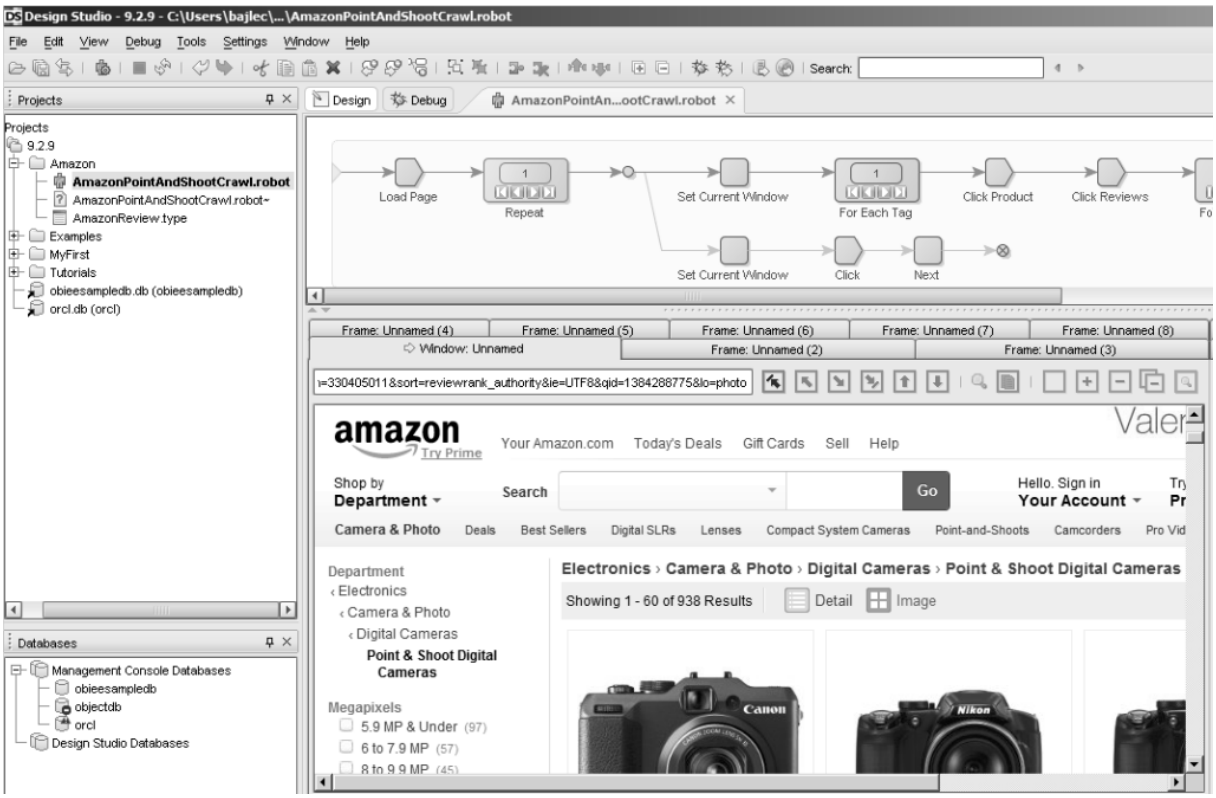

Abb. 4: Schnittstellen- bzw. Robot-Design mit Kapow Katalyst

Sobald die Datenintegrationsabläufe erstellt sind, werden sie auf die Kapow Katalyst Management Konsole hochgeladen, siehe Abb. 5. Hier können sie so programmiert werden, dass sie Arbeitsprozesse entweder terminiert gebündelt oder on-demand als REST-Service ausführen.

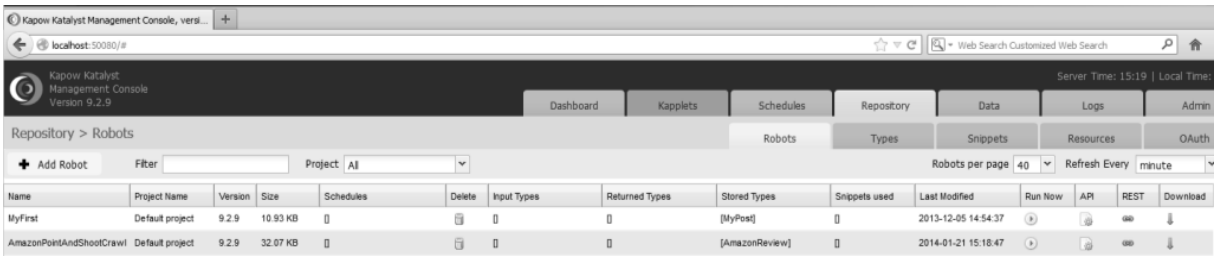

Abb. 5: Robot-Automation in der Kapow Management Console

In der Management Konsole werden die Funktionen der Kapow Enterprise Plattform verwaltet. Eine rollenbasierte Administration unterstützt die sichere, granulare Kontrolle der Integrations-Projekte, der Zugangsrechte auf die Kapplets selbst und die erzielten Ergebnisse. Der Zugang zu den Kapplets kann über LDAP oder Active Directory sowie andere Directories festgelegt werden oder über die integrierten Benutzermanagementfunktionen erfolgen. Die Management Konsole arbeitet auch mit Integrationsabläufen in Form REST/SOAP-basierter Web Services, Flows zur Integration von Daten, die in Dokumente, E-Mail, SQL oder NoSQL Datenbanken oder Hadoop Umgebungenn überführt werden sollen, und kreiert automatisch die nötigen APIs zur Nutzung von Java-, .NET- oder BI-Lösungen.

Im Ergebnis wird auf diesem Weg eine Synthetic API bereitgestellt und damit das Integrationsproblem gelöst: von einer Web Applikation bzw. einem Portal, das nicht über eine API verfügt, lassen sich nun Daten und Services extrahieren bzw. integrieren, das Prinzip zeigt Abb. 6.

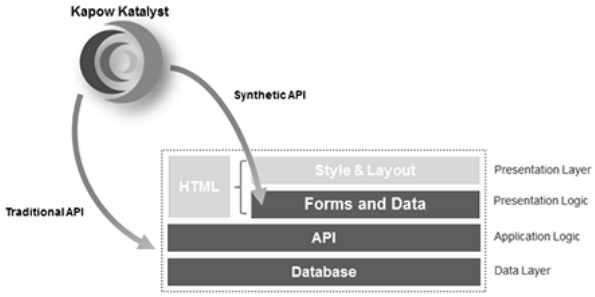

Abb. 6: Agile Datenintegration unabhängig von APIs und Konnektoren

Data Mashup mit Endeca Information Discovery

Mit Endeca Information Discovery (EID) lassen sich schließlich Anwendungen für explorative Analysen erstellen, konsumieren und unternehmensweit teilen. Die intuitive Benutzeroberfläche erlaubt eine selbstständige Datenerkundung über alle am Endeca Server angebundenen Datenquellen hinweg [5] [6]. Darüber hinaus bietet EID Studio ab Version 3.1 auch eine Möglichkeit zum Laden von JSON- oder MS Excel-Dateien oder auch den direkten JDBC-Zugriff auf freigegebene Datenbanken, wie z.B. auf ein Oracle Data Warehouse mit den Kamera-Verkaufszahlen aus dem oben beschriebenen Szenario bzw. auf die mit Kapow Katalyst erfassten Kundenrezensionen von Amazon.com.

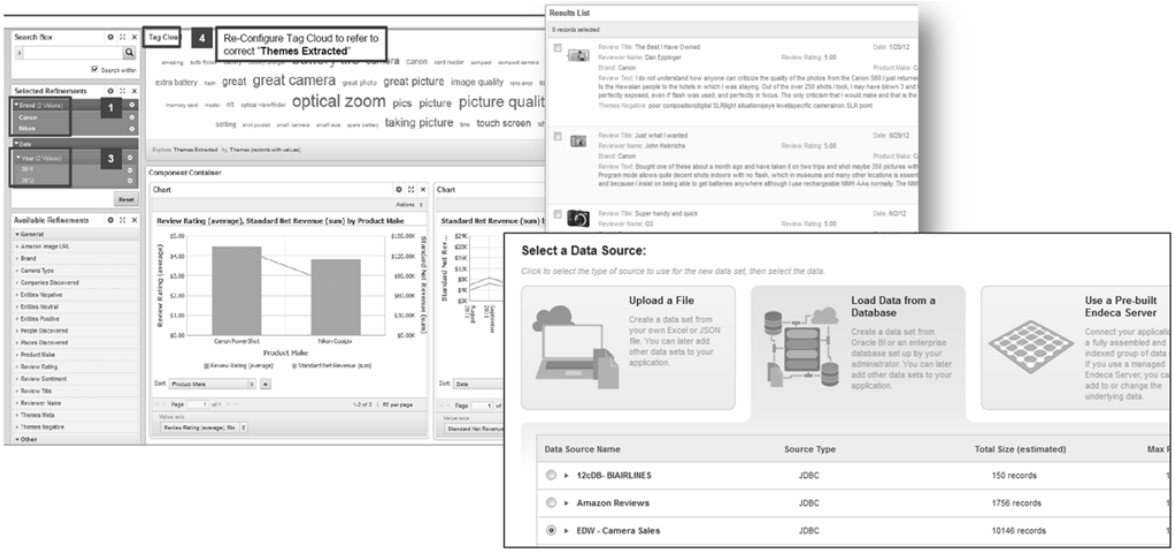

Abb. 7: Kombinierte Analyse von Kameraverkäufen vs. Kundenrezensionen mit Endeca Information Discovery

Weiterführende Informationen

[1] Oracle White Paper - Information Management and Big Data, A Reference Architecture, 2014 → Link

[2] Jon Mead, Marc Rittman, Stewart Bryson, Andrew Bond: Introducing the Update Oracle / Rittman Mead Information Management Architecture

Part 1 – Information Architecture and the "Data Factory" → Link

Part 2 – Delivering the Data Factory → Link

[3] Stewart Bryson, Andrew Bond: Evolution of Information Management - Architecture and Development → Link

[4] Endeca Information Discovery Integrator 3.1.x Documentation: Web Acquisition Toolkit (WAT) 3.1.1 → Link

[5] Helen Sun, William Smith: Master Competitive Analytics with Oracle Endeca Information Discovery (Oracle Press, 2014) → Link

[6] Harald Erb: "Model as you go" & Data Mashup bei der Datenerkundung mit Oracle Endeca, Manuskript DOAG 2013 Konferenz → Link

[7] Tom Plunkett, Brian Macdonald, Bruce Nelson, Mark Hornick, Khader Mohiuddin und andere: Oracle Big Data Handbook (Oracle Press, 2013) → Link

Kontakt:

Harald Erb	Telefon:	+49 (0) 6103-397 403
ORACLE Deutschland B.V. & Co. KG	Fax:	+49 (0) 6103-397 397
Robert-Bosch-Straße 5	E-Mail	harald.erb@oracle.com
D-63303 Dreieich	Internet:	www.oracle.com