Nr. 4 | August 2012 | ISSN 09 36-0360

DOAG News

Deutsche ORACLE-Anwendergruppe e.V.

Data Warehouse & BI

www.doag.org

OPITZ CONSULTING

Menschen.
Innovationen.
Lösungen.




Wenn Ihnen große Datenmengen schwer zu schaffen machen ...

... dann ist unsere Lösung für Sie das All-in-one-System Oracle Exadata!

Testen Sie jetzt die Exadata mit Ihren eigenen Daten auf Ihrer eigenen Infrastruktur!

Wir bieten Ihnen den Proof of Concept ohne „Try & Buy"-Bedingungen – Sie entscheiden sich allein aufgrund Ihrer Analysen für oder gegen die Exadata!

Sie möchten Ihre Datenbanklandschaft konsolidieren? Sie benötigen eine hoch performante Datenbankumgebung? Sie müssen große Datenmengen effektiv verarbeiten?

Nutzen Sie bewährte und auch neue Features wie

- Exadata Smart Scan,
- Exadata Smart Flash Cache,
- oder Exadata Hybrid Columnar-Komprimierung

kombiniert und integriert in einem Engineered System – Oracle Exadata.

Sie möchten sicher gehen, dass Exadata für Ihr Unternehmen die richtige Lösung bereitstellt? Ihnen fehlt das Know-how oder die Zeit für die Einführung und Umsetzung eines Exadata-Projektes?

- Wir beraten Sie bei der Planung, unterstützen Sie aktiv in einem Proof of Concept und bei der Einführung von Exadata.
- Sie testen Ihre Anwendung mit Ihren Daten in Ihrem Rechenzentrum, wir kümmern uns um den Rest!
- Auf Wunsch übernehmen wir den gesamten Lifecycle Ihres Oracle Exadata Projektes:
  Von der Planung, Implementierung und Schulung bis zum Betrieb mit unserer Remote IT-Administration (PASM®) für Exadata Service.

Ihr direkter Ansprechpartner zu unserem Exadata Leistungsangebot ist Chris Bochow.
Telefon: +49 2261 6001-0 · E-Mail: chris.bochow@opitz-consulting.com

Mehr zum Thema:

**Oracle Exadata-
Plan, Build, Run!**

Lesen Sie mehr zu unseren Leistungen für Exadata Plan, Build und Run.

www.opitz-consulting.com/exadata

Sie suchen eine kleinere Lösung?

Die Oracle Database Appliance (ODA) bietet Ihnen Server, Storage, Netzwerk aus einer Hand. Zusammen mit der DB 11gR2 leistet das voll integrierte System alles, was Sie zum Betrieb von selbst entwickelten, fertigen OLTP Anwendungen oder Data Warehouses brauchen.

www.opitz-consulting.com/oda

ORACLE Platinum Partner

Christian Weinberger
Leiter Bereich Data Warehouse und Business Intelligence

Liebe Mitglieder der DOAG Deutsche ORACLE-Anwendergruppe,
liebe Leserinnen und Leser,

Sie können das Thema „Big Data“ langsam nicht mehr hören und sehen darin keinen Nutzen für Ihr Unternehmen? Ich verstehe Sie, zumindest ein wenig! Lange gab es kein vergleichbares Hype-Thema mehr im BI-Bereich. Die unterschiedlichen Sichtweisen und Interpretationen der Hersteller tun ein Übriges: Von „VVVV“ über „VVV“ bis hin zu „einfach große Datenvolumina“ ist im Markt fast jede Deutung zu finden. Fakt ist, dass „Big Data“-Technologien im ursprünglichen Sinne derzeit eine Speziallösung für Anwender mit sehr großem Datenvolumen sind. „Sehr groß“ beginnt hier im oberen zweistelligen Terabyte-Bereich und geht bis in die Petabytes. Ausnahmen bestätigen wie immer die Regel. Trotzdem werden wir in den kommenden Jahren beobachten können, wie klassische Big-Data-Prinzipien, wie auf Dateisystem-Ebene verteilte Storages oder Verarbeitungen nach dem „Map & Reduce“-Prinzip, auch in den Mainstream Einzug halten und alltagstauglich für die breite Masse der Anwender werden.

In anderen Bereichen erleben wir dafür Veränderungen, die schon in naher Zukunft viele von uns tangieren werden: Agile BI und Selfservice-Ansätze führen zu einer kleinen organisatorischen Revolution und letztlich zu einer neuen Definition der Zusammenarbeit zwischen Anwendern und BI-Stäben beziehungsweise BICCs. Aber auch die technologischen Basics vieler aktueller Data-Warehouse-Installationen erfahren durch technische und systemische Innovationen einen deutlichen Schub nach vorn. Automatisierungsansätze und Frameworks versprechen Effizienz-Gewinne für bestehende Infrastrukturen und die Oracle-Zukäufe aus jüngerer Zeit zeigen hier ganz neue Wege auf.

Die DOAG unterstützt Sie in Ihren BI- und DWH-Vorhaben zum einen mit dieser Publikation und zum anderen mit einem umfangreichen Angebot an Veranstaltungen, Vorträgen und Seminaren. In diesem Sinne wünsche ich Ihnen viel Spaß bei der Lektüre der aktuellen DOAG *News*.

ORACLE Platinum Partner

HUNKLER GmbH & Co. KG

„Best Solutions based on Oracle, von einem der führenden Oracle-Systemhäuser in Deutschland“

LIZENZBERATUNG & -VERTRIEB

HOCHVERFÜGBARKEITSLÖSUNGEN & PERFORMANCE TUNING

DATA WAREHOUSING & BUSINESS INTELLIGENCE LÖSUNGEN

ORACLE APPLIANCES

HUNKLER – die erste Adresse beim Thema Oracle

Ausfallsichere Datenbanken, professionelle Lösungen für Business Intelligence, leistungsstarke Appliances: Auf diese Schwerpunkte haben wir uns nach den von Oracle vorgegebenen Anforderungen spezialisiert. Spezialisten für Oracle sind wir schon seit 1987, als wir erster offizieller Partner in Deutschland wurden.

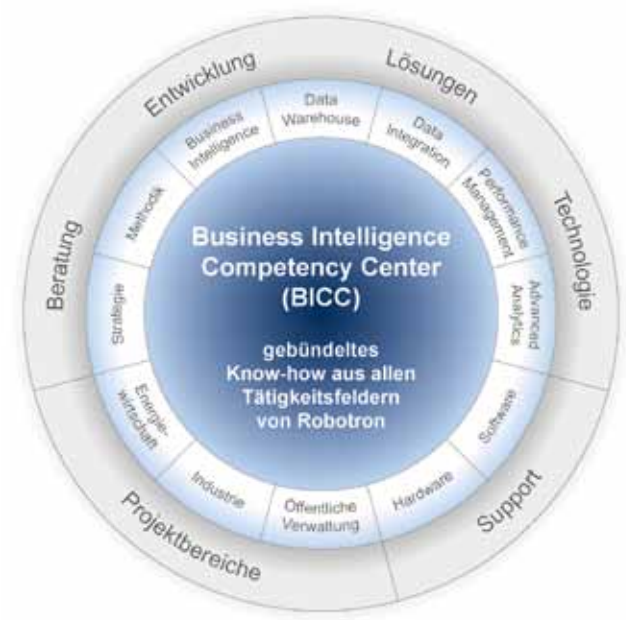
Wir wissen genau, was der Mittelstand wirklich braucht: modernste Technologie, zugeschnitten auf individuelle Business-Lösungen, die sofort Kosten senken. Lösungen, mit denen Unternehmen von Anfang an spürbare Wettbewerbsvorteile erzielen und langfristig festigen können.

Von der Systemplanung bis zum Lizenzmanagement. Es gibt immer den richtigen Weg zu mehr Effizienz in der IT. Bei uns. Für Sie.

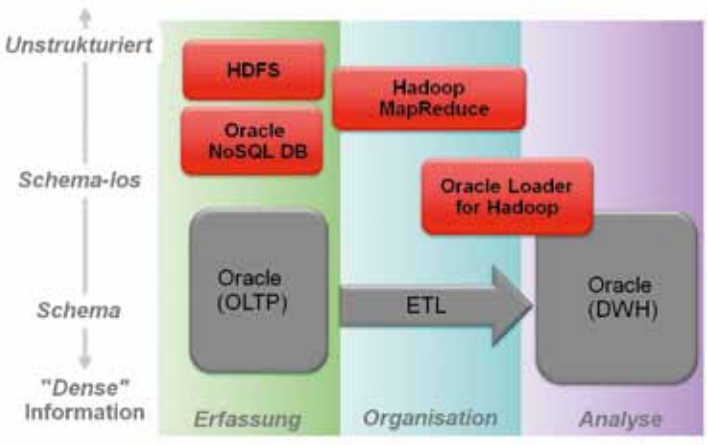
Hauptsitz Karlsruhe
Bannwaldallee 32, 76185 Karlsruhe, Tel. 0721-490 16-0, Fax 0721-490 16-29
info@hunkler.de, www.hunkler.de

Geschäftsstelle Bodensee
Fritz-Reichle-Ring 6a, 78315 Radolfzell, Tel. 07732-939 14-00, Fax 07732-939 14-04
info@hunkler.de, www.hunkler.de

DWH/BI-Framework und Vorgehensmodell: der Weg zur BI Excellence, Seite 24



To BICC or not to be – auch für einen IT-Dienstleister, Seite 30



Einführung für RDBMS-Kenner: Hadoop, MapReduce und mehr, Seite 51

Spotlight

Mittwoch, 9. Mai 2012

Dr. Frank Schönthaler, Leiter der DOAG Business Solutions Community, überreicht im Rahmen der DOAG 2012 Applications die Auszeichnung „DOAG Botschafter Applications“ an Matthias Forkel, IT-Leiter des mittelständischen Unternehmens Prodinger Verpackungen in Coburg. Den Preis erhalten Personen, die sich durch ihr hohes Engagement für die DOAG ausgezeichnet haben.

Dienstag, 5. Juni 2012

Die DOAG 2012 Logistik + SCM findet mitten in Hamburgs historischer Speicherstadt im ehemaligen Hauptzollamt statt. Die Business Solutions Community der DOAG fokussiert dort Logistik-Lösungen auf der Plattform Oracle sowie den Erfahrungsaustausch unter Kollegen und bietet den rund 80 Teilnehmern ein vielseitiges Vortragsprogramm, prominente Keynote-Speaker und praxisnahe Workshops. Erfreulich ist die deutlich gestiegene Zahl der Endanwendervorträge.

Dienstag, 19. Juni 2012

Dr. Dietmar Neugebauer stellt auf dem Jahreskongress der Austrian Oracle User Group (AOUG) die DOAG vor und erläutert das Potenzial einer neuen Zusammenarbeit.

Freitag, 22. Juni 2012

Die DOAG-Leitung verabschiedet einstimmig die neue Satzung der DOAG.

Montag, 25. Juni 2012

Dr. Dietmar Neugebauer, Vorstandsvorsitzender der DOAG, und Fried Saacke, DOAG-Vorstand und Geschäftsführer, stellen abwechselnd die neue Satzung der DOAG deutschlandweit auf allen Regionaltreffen vor. Auftakt ist in Würzburg, wo der innovative Vorschlag sehr positiv aufgenommen wird.

Donnerstag, 5. Juli 2012

Wolfgang Taschner, Chefredakteur der Java aktuell, Dr. Dietmar Neugebauer und Fried Saacke präsentieren den Interessenverbund der Java User Groups e.V. (iJUG), bei dem die DOAG Mitglied ist, auf dem Java Forum in Stuttgart. Das Interesse der 1.200 Besucher an der Zeitschrift ist groß.

Freitag, 6. Juli 2012

Dr. Dietmar Neugebauer und Fried Saacke treffen sich in Walldorf mit Dr. Mario Günter, Geschäftsführer der Deutschsprachigen SAP-Anwendergruppe e.V. (DSAG), zu einem Interview über die Arbeit der Anwendergruppen, das in der nächsten Ausgabe der DOAG News und der DSAG-Blaupause gemeinsam veröffentlicht wird.

Montag, 9. Juli 2012

Bei einem Treffen mit Jürgen Kunz, Geschäftsführer der ORACLE Deutschland B.V. & Co. KG, besprechen Dr. Dietmar Neugebauer und Fried Saacke die Optimierung der Zusammenarbeit. Ziel ist es, Probleme an die richtigen Personen bei Oracle zu adressieren, um eine schnelle Lösung zu erhalten.

Freitag, 20. Juli 2012

Die DOAG-Leitung und die Verantwortlichen der einzelnen Streams finalisieren in München das Programm der DOAG 2012 Konferenz + Ausstellung. Vom 20. bis zum 22. November 2012 erwartet die Teilnehmer im Nürnberg ConventionCenter wieder ein umfangreiches und qualitativ hochwertiges Programm mit mehr als 400 Vorträgen und vielen Highlights aus der Datenbank, der Development und der Infrastruktur Community.

Dienstag, 26. Juni 2012

Dr. Dietmar Neugebauer und Fried Saacke gratulieren der Swiss Oracle User Group (SOUG) zum 25-jährigen Bestehen und überreichen ein Geschenk der DOAG.

Die DOAG in der Presse

Die nachfolgenden Ausschnitte reflektieren die Einschätzung der Fach- und Wirtschaftspresse zu bestimmten Themen über Oracle; die Veröffentlichungen geben nicht die Meinung der DOAG wieder und sind auch nicht im Vorfeld mit der DOAG abgestimmt. Lediglich die Zitate einzelner DOAG-Vorstände geben die Meinung der DOAG wider.

Computerwoche vom 14. Mai 2012

Sicherheitslücke nervt Oracle-Kunden

Obwohl die Sicherheitslücke „TNS Poison“ in der Oracle-Datenbank seit vier Jahren bekannt ist, gibt es bis heute keinen Patch. Anwender müssen selbst Hand anlegen.

„Es ist ein Armutszeugnis für Oracle“, kritisieren Vertreter der Deutschen Oracle Anwendergruppe (DOAG). Im Jahr 2008 hatte der Security-Spezialist Joxean Köret eine nach seinen Worten schwerwiegende Sicherheitslücke (TNS Poison) in Oracles Datenbanken entdeckt und den Hersteller darüber informiert. In der Annahme, Oracle habe das Problem behoben, veröffentlichte Köret am 18. April in einem Blog-Eintrag weitergehende Informationen – vor allem auch, um die Anwender dazu zu bewegen, die aktuellen Patches für ihre Datenbanken einzuspielen …

… Oracle reagierte – allerdings nicht mit einem Patch, sondern mit einem manuellen Workaround, was laut DOAG unter den Kunden großes Staunen hervorgerufen habe. Diese müssten durch das Verändern von Parametern die Lücke schließen. Für Real-Application-Cluster-(RAC-)Umgebungen empfiehlt Oracle eine SSI7TLS-Verschlüsselung. Diese Option, die sich Oracle normalerweise bezahlen lässt, könnten betroffene Kunden nun kostenlos nutzen. Ein Zeichen dafür, dass der Hersteller das Problem offenbar als gravierend einschätzt.

Security-Spezialist und DOAG-Vertreter Alexander Kornbrust von der Firma Red-Database-Security GmbH vermutet, da es nun einen Workaround gebe, werde Oracle so schnell wohl keinen Patch für das Problem herausbringen.

ix vom 18. Juni 2012

Oracle antwortet auf Kundenkritik

„Alles halb so schlimm“, so lautet der Tenor der Antwort, mit der Oracle auf Kritik der deutschen Anwendervereinigung DOAG an Version 3 der Virtualiserungstechnik VM reagiert. Der Verein hatte unter anderem Schwierigkeiten bei Upgrades und fehlende Verwaltungstools bemängelt. In seinem Schreiben verweist der Hersteller auf Erfahrungen anderer Kunden sowie die Software Enterprise Manager, die die benötigten Funktionen mitbringe.
Die DOAG freut sich zwar über die schnelle Reaktion, zufrieden ist sie damit jedoch nicht. Zwar seien einige der Mängel mit der aktuellen Release VM 3.1.1 behoben. Letztlich sei das Produkt aber zu umständlich zu administrieren: Neben dem VM Manager gebe es mit dem Enterprise Manager 12c und dem Enterprise Manager Ops Center zwei weitere grafische Verwaltungswerkzeuge. Alle drei böten unterschiedliche Funktionen und benötigten ihre eigene Instanz des Application-Servers Weblogic, gegebenenfalls sogar noch ein Datenbank-Repository.
Das sei möglicherweise für große Firmen akzeptabel, Anwender in kleineren Organisationen hätten jedoch andere Bedürfnisse. Kleine Firmen bräuchten flexible, leicht einsetzbare Lösungen, schreibt die DOAG. „Im Kontext der Hochverfügbarkeit ist das Konzept von Oracle für kleinere virtuelle Umgebungen sogar unwirtschaftlich: Man braucht sechs bis acht physische Maschinen, um ein paar virtuelle Server zu betreiben. Es liegt auf der Hand, dass dies nicht praktikabel ist“, sagt Björn Bröhl, Leiter der Infrastruktur & Middleware Community bei der DOAG.

ZDNet vom 4. Juli 2012

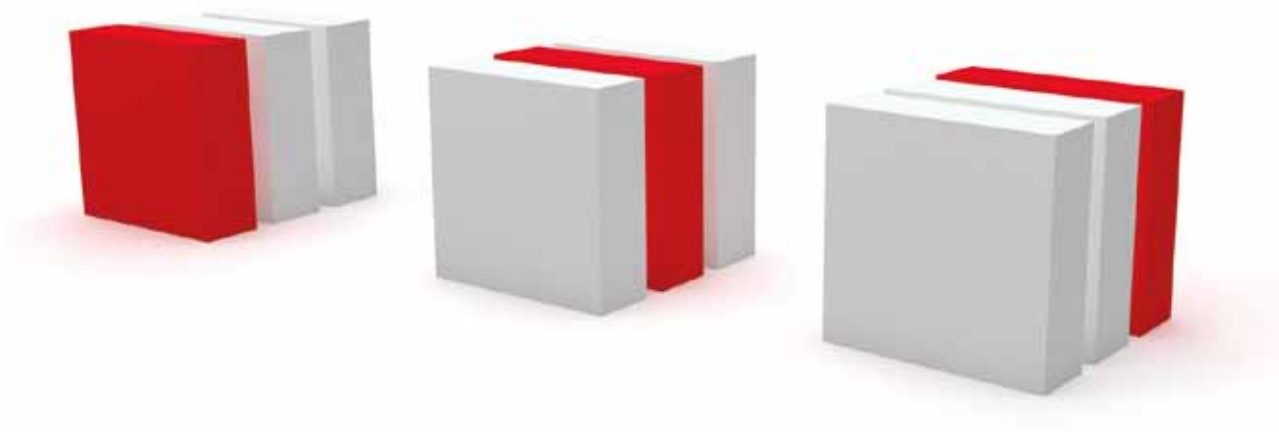
Oracle-Kunden freuen sich über Urteil des EuGH zu Gebrauchtsoftware

Die Deutsche Oracle-Anwendergruppe e.V. (DOAG) hat sich jetzt zum Urteil des Europäischen Gerichtshofs im Streit zwischen dem Hersteller und UsedSoft geäußert. „Wir begrüßen das Urteil des Europäischen Gerichtshofs, da dies die Investitionssicherheit der Kunden stärkt”, teilt DOAG-Vorstandsvorsitzender Dietmar Neugebauer in einer Presseausendung mit. „Sollte der Bundesgerichtshof diese Entscheidung bestätigen, würde es zu einer Liberalisierung des Marktes führen, die im Sinne der Anwender ist.”

Das Gericht in Luxemburg hatte gestern auf Ersuchen des Bundesgerichtshofs entschieden, dass gebrauchte Softwarelizenzen unabhängig davon weiterverkauft werden dürfen, ob der Erstkäufer sie ursprünglich per Download oder auf einem Datenträger erworben hat.

Weitere Pressestimmen lesen Sie unter http://www.doag.org/presse/spiegel

BI-Performance by Design



Strukturiert. Konsequent. Nachhaltig.

Erfolg und Akzeptanz eines Business Intelligence Systems hängen von der Performance ab. Near Time Reporting und neue Nutzergewohnheiten wie Mobile BI erhöhen die Anforderungen an die Systeme.
Von essentieller Bedeutung sind:

- eine klar strukturierte Datenmodellierung
- die konsequente Nutzung der Technologien
- die Nachhaltigkeit durch Projektstandards

Noch bis zum 30. September!
Ihr direkter Kontakt
02 21-66 95 75-75

Jetzt kostenlose thematische Einführung bei Ihnen vor Ort anfordern!
performance@areto-consulting.de

ORACLE Gold Partner

www.areto-consulting.de

ARETO consulting

Fotos: Wolfgang Taschner

Die optimale Nutzung der Datenbanken ist die Basis für das Geschäft der Postbank. Christian Trieb, Leiter der DOAG Datenbank Community, und Wolfgang Taschner, Chefredakteur der DOAG *News*, sprachen darüber mit Jens-Christian Pokolm (rechts), zuständig für Datenbank-Design bei der Postbank Systems AG.

„Die Stabilität bestehender Software ist mir wichtiger als neue Funktionen …“

Was sind die größten IT-Herausforderungen bei der Postbank?

Pokolm: Die größte Herausforderung ist der Kostendruck. Wir bieten als Bank unseren Kunden ein kostenloses Girokonto an, das heißt, wir bekommen kein Geld für Zahlungsoperationen wie beispielsweise beim Online-Banking. Trotzdem muss der Prozess in der IT abgebildet sein. Hinzu kommt eine Höchstverfügbarkeit der Systeme rund um die Uhr. Zudem erwartet der Kunde immer die neuesten Features hinsichtlich Bedienbarkeit und Funktionalität.

Wie lösen Sie die Hochverfügbarkeit in der Praxis?

Pokolm: Wir betreiben grundsätzlich ein Active-Active-Cluster und keine Single-Instance-Lösung mehr. Damit existiert für den Datenbank-Administrator auf den verschiedenen Plattformen nur noch eine einheitliche System-Umgebung. Das Ganze findet unter einem sogenannten „Short-Stretched-Ansatz“ über mehrere Rechenzentren verteilt statt, um Wartung, Upgrades etc. fahren zu können. Dabei halten wir den Software-Stack möglichst gering, um Fehlerquellen durch unterschiedliche Anbieter auszuschalten.

Wie ist der Weg hin zur Hochverfügbarkeit verlaufen?

Pokolm: Da mussten wir natürlich einige schmerzhafte Erfahrungen machen. Zum Glück haben wir im Bereich „Engineering“ ein großes und motiviertes Team, sodass wir viele Dinge selbst intern testen konnten. Zudem sind wir über die einzelnen Releases, angefangen bei Oracle Parallel Server, Schritt für Schritt in die jetzige Umgebung gewachsen, wobei der große Durchbruch mit der Datenbank-Version 10 Release 2 erfolgte. Aufgrund des durchgängigen Stacks sind damit die Fehlermöglichkeiten deutlich weniger geworden.

Welchen Umfang hat der Oracle-Stack bei Ihnen?

Pokolm: Er beginnt heute oberhalb des Betriebssystems und umfasst neben der Enterprise Edition und RAC einige wenige Optionen. Künftig wollen wir zusätzlich auch auf Oracle Enterprise Linux setzen. Da Linux eine offene Plattform ist, sind wir hier nicht ausschließlich auf Oracle angewiesen.

Welche Erfahrungen hinsichtlich Hochverfügbarkeit können Sie an andere Unternehmen weitergeben?

Pokolm: Die größte Herausforderung besteht darin, einen sauberen Hardware- und Betriebssystem-Stack zu finden und einzurichten, um eine sichere

Infrastruktur aufsetzen zu können. Die andere wichtige Voraussetzung ist gut ausgebildetes Personal mit dem richtigen Verständnis für die Problematik.

Welchen Umfang hat Ihre Hochverfügbarkeitslösung?

Pokolm: In der produktiven Umgebung haben wir rund vierhundert Datenbanken. Daneben gibt es eine exakte Kopie der Produktion zum Test, sodass wir dort die gleichen Bedingungen haben wie in der realen Umgebung, sowie eine weitere Kopie für die Entwicklung. Die Kapazität der einzelnen Datenbanken reicht von wenigen Gigabyte bis in den zweistelligen Terabyte-Bereich. Zur Administration sind zehn DBAs im Einsatz.

Wie ist das Verhältnis zwischen Standard-Software und Eigenentwicklung?

Pokolm: Da wir gemeinsam mit der SAP das Core-Banking entwickeln, sind bei uns relativ viele SAP-Entwickler beschäftigt. Die gesamte Banking-Middleware ist eine komplette Eigenentwicklung.

Wie halten Sie Ihre Datenbanken auf dem aktuellen Stand?

Pokolm: Da Oracle seine Patchsets mittlerweile ankündigt beziehungsweise regelmäßig durchführt, ist die Situation einfacher geworden. Das große Problem ist, dass nicht alle Patches online eingespielt werden können. Andererseits müssen wir aus Revisionsgründen die Patches zeitnah ausführen.

Wie garantieren Sie die Sicherheit Ihrer Daten?

Pokolm: Das Thema „Security" ist einfach zu handhaben, da unsere Daten relativ gut nach außen abgeschottet sind. Unsere Grundregel lautet „Wo kein Kabel, da kein Risiko". Das reine Produktionsnetz besitzt keine direkte Verbindung zum Internet, um bestimmte Sicherheitsrisiken von vornherein auszuschließen. Auch beim Online-Banking gibt es eine harte Entkopplung zwischen Internet und Bank. Hinzu kommen weitere Sicherheitsmaßnahmen wie Firewalls, Sensoren sowie das zyklische Patchen im Security-Kontext.

Wie gehen Sie dann mit den Online-Tools von Oracle-Support um?

Pokolm: Patches in unserer Produktions-Umgebung können nur auf einem physischen Medium wie einem USB-Stick oder einer CD nach einem entsprechenden Scan eingespielt werden. Auch umgekehrt übertragen wir keine Systems-Daten online an Oracle.

Wie rollen Sie Patches auf mehrere Hundert Datenbanken aus?

Pokolm: Wir setzen dazu den Oracle Enterprise Manager ein. Hinzu kommen entsprechende interne Werkzeuge. Auf den hochkritischen Systemen gibt es aber ganz klar die Regel, dass alles, was zu Geldschäden führen könnte, von Hand ausgeführt werden muss, um im Fehlerfall sofort reagieren zu können. Darüber hinaus werden alle Patches zuvor in der Test- und in der Entwicklungs-Umgebung intensiv ausprobiert.

Wie beurteilen Sie die Zukauf-Strategie von Oracle?

Pokolm: Ich habe hier gemischte Gefühle. Zum einen ist es natürlich ein Vorteil, wenn ein Hersteller den gesamten Stack anbieten kann. Problematisch ist jedoch, dass man dann einem Monopolisten ausgeliefert ist, der beispielsweise die Supportgebühren für die Hardware plötzlich um Faktoren erhöht. Da ist auch die DOAG eine gute Anlaufstelle, um die Interessen der Kunden gegenüber dem Hersteller zu vertreten.

Ist Exadata ein Thema für die Postbank?

Pokolm: Stand heute brauchen wir Exadata noch nicht. Wir haben uns allerdings bereits damit beschäftigt. Exadata ist schwierig in eine Rechenzentrums-Infrastruktur einzupassen, außerdem kostet das System eine ganze Menge.

Welche Wünsche haben Sie an Oracle?

Zur Person: Jens-Christian Pokolm

Nach der Ausbildung zum Datenverarbeitungskaufmann (IHK) und dem Studium der Elektrotechnik begann Jens-Christian Pokolm 1988 als Organisations-Programmierer. Weitere Stationen seiner Karriere waren Software-Entwicklungsleiter für „CD-Search", einer Volltext-Retrieval Datenbank, Projektleiter bei Informix und Projektleiter für „Core- and New-Technologie" im Bereich öffentliche Verwaltung bei Oracle Deutschland in Bonn. Er arbeitet seit 1999 bei der Postbank Systems AG in verschiedenen Funktionen, alle im Datenbank-Kontext: Projektleitung, Betriebs- und Systemplanung, Evaluation und Einführung neuer Technologien, Projektplanung- und Beratung, Architektur und Design sowie Richtlinien und Konzepte für Datenbanken. Jens-Christian Pokolm ist Oracle Certified Professional, Autor des Buchs „Oracle Apex und Oracle XE in der Praxis" und hält regelmäßig Vorträge, unter anderem auf der DOAG Konferenz + Ausstellung und der Oracle OpenWorld.

Pokolm: Ich wünsche mir, dass Oracle weniger „Bananen-Produkte" produziert, die grün zum Kunden geliefert werden, um dort zu reifen. Für mich muss ein System auf einer gängigen Umgebung nach der Installation laufen und nicht erst nach einigen Patches – auch wenn es sich um eine erste Version handelt. Oracle arbeitet nach meinem Empfinden daran, doch das

Das Unternehmen

Die Postbank Systems AG ist der IT-Dienstleister der Deutsche Postbank AG. Das Leistungsangebot umfasst alle Produkte des IT-Betriebs sowie alle IT-Projekte. Postbank Systems beschäftigt rund 1350 Mitarbeiter. Die Qualitäts- und Kostenführerschaft bei Standardprodukten wird über die Realisierung von Skaleneffekten und durch Standardisierung und Automatisierung der Prozesse weiter ausgebaut. Dabei werden konsequent die Chancen von zukunftssicheren und wettbewerbsfähigen Technologien genutzt.
Die Postbank ist eine der führenden Multikanalbanken in Deutschland. Postbank und Postbank Systems setzen auf den Ausbau von integrierten Multikanalservices. Im Direktbankbereich stellt die Postbank Systems innovative und effiziente Lösungen kostengünstig zur Verfügung. Hoch verfügbare und leicht verständliche Direktservices machen Bankgeschäfte bequem und einfach. Die führende Position wird auf Basis von State-of-the-art Technologien ausgebaut.
Ausblick: Postbank Systems unterstützt mit ihrer konsequenten Ausrichtung auf Standardisierung, Effizienz und Massenfähigkeit die strategische Ausrichtung der Postbank auf das Retailgeschäft.

könnte deutlich schneller gehen. Die Stabilität bestehender Software ist mir wichtiger als neue Funktionen. Jeder aufgetretene Datenbank-Fehler bedeutet mehr Ärger als ein nicht vorhandenes Feature.

In welche Richtung wird sich Ihre IT in den kommenden Jahren entwickeln?

Pokolm: Was sich ändern wird, ist die Hardware-Architektur, speziell hinsichtlich des Themas „Private Cloud". Außerdem sollte es möglich sein, die Hardware besser auszunutzen. Bei der Datenbank werden sich sicher die sogenannten „Self-Services" durchsetzen.

Wie sehen Sie den Stellenwert einer Anwendergruppe wie der DOAG?

Pokolm: Gerade unsere jungen DBAs lernen sehr viel auf den DOAG-Veranstaltungen.

Impressum

Herausgeber:
DOAG Deutsche ORACLE-Anwendergruppe e.V.
Tempelhofer Weg 64, 12347 Berlin
Tel.: 0700 11 36 24 38
www.doag.org

Verlag:
DOAG Dienstleistungen GmbH
Fried Saacke, Geschäftsführer
info@doag-dienstleistungen.de

Chefredakteur (ViSdP):
Wolfgang Taschner, redaktion@doag.org

Redaktion:
Fried Saacke, Carmen Al-Youssef, Mylène Diacquenod, Christian Weinberger, Stefan Kinnen, Dr. Frank Schönthaler, Christian Trieb

Titel, Gestaltung und Satz:
Claudia Wagner, Fana-Lamielle Samatin
DOAG Dienstleistungen GmbH

Titelfoto: Fotolia

Anzeigen:
CrossMarketeam Doris Budwill
www.crossmarketeam.de
Mediadaten und Preise finden Sie unter:
www.doag.org/go/mediadaten

Druck:
adame Advertising and Media GmbH Berlin, www.adame.de

Data-Warehouse-Systeme gehören heute zum Unternehmensalltag. Sie sind strategisch und oft mit wichtigen operativen Anwendungen verzahnt. Aber in kaum einem anderen Bereich gibt es so viele Baustellen, unzufriedene Anwender und klagende Kostenträger.

Konzepte und Methoden im Oracle Data Warehouse

Alfred Schlaucher, ORACLE Deutschland B.V. & Co. KG

Die größten Probleme beim Betrieb von Data Warehouses sind: zu geringe Umsetzungsgeschwindigkeit neuer Analyse-Anforderungen, Komplexität und Kosten für Betrieb und Weiterentwicklung, zu geringe Informationsausbeute für die Benutzer und zu wenig Anwenderkomfort hinsichtlich Performance und Flexibilität. Auf der anderen Seite boomt der Markt für Tools und es lösen sich die Wellen vermeintlicher Innovationen gegenseitig ab: Appliance, Realtime Data Warehouse, In-Memory, Big Data etc. Doch die Ursachen der Probleme sind nicht schlechte Techniken oder Tools, oft werden wichtige Methoden und Architekturgrundsätze nicht angewandt. Deshalb an dieser Stelle eine Zusammenfassung von einfachen Regeln und Methoden, um Anwendung und Betrieb des Oracle Data Warehouse effizienter zu machen – ohne dass es teuer wird.

Theoretische Grundlagen

Seit Einführung des Data-Warehouse-Konzepts in den 1990er Jahren gelten vier Prinzipien für ein erfolgreiches Data Warehouse (DWH):

- Unternehmensweite DWHs halten Informationen an zentraler Stelle in einer integrierten und abgestimmten Form vor
- Die Informationen sind für alle verständlich abgelegt, ohne Fachjargon und mit zusätzlichen fachlichen Erklärungen und Referenzinformationen
- Die Daten sind historisiert, was Vorhersagen für die Zukunft erlaubt
- Die Daten sind aus den OLTP-Systemen herauskopiert und ermöglichen eine neutrale, von operativen Zwängen losgelöste und Zeitpunktbezogene Analyse

Damit OLTP-Daten diesen Zustand erreichen, durchlaufen sie in dem DWH einen notwendigen Informationsbeschaffungsweg, den man in drei Phasen (Schichten / Layer) unterteilen kann:

- Herauslösen der Daten aus den Vorsystemen (OLTP), Integrieren, Harmonisieren und Qualitätssichern (Stage-Schritt, Data Integration Layer)
- Granularisiertes Ablegen der Informationen (nahezu drei NF), Anreichern mit Referenzdaten und Historisieren (Enterprise Information Layer)
- Endbenutzergerechtes, fachthemenorientiertes Aufbereiten (meist multidimensional organisierte Daten, User View Layer)

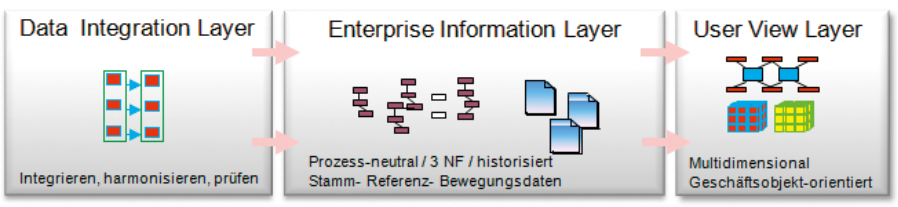


Abbildung 1: Klassisches Drei-Schichten-Modell einer DWH-Architektur

Dieses Drei-Schichten-Modell hat sich als Standard etabliert (siehe Abbildung 1). Auf das Oracle Data Warehouse angewendet gelten nachfolgend beschriebene Konzepte beziehungsweise Best Practices.

Konzepte, Methoden und Best Practices

Die Konzentration des Schichtenmodells auf eine einheitliche, zusammenhängende physische Ablage ist gerade für das Oracle Data Warehouse eine der wichtigsten Forderungen und zielt auf eine Reihe von Vorteilen, die nachfolgend beschrieben sind. Alle oben genannten Layer sind am besten in einer Datenbank auf einem Server (oder Server-Cluster) aufgehoben.

Ein wichtiges Prinzip ist das datenzentrierte Vorgehen beim Management der Warehouse-Daten. Das beinhaltet das Nutzen der technischen Mittel der Datenbank als primärer Host der Daten und betrifft Security, ETL, Vorbereiten von Kennzahlen und auch fachspezifische Sichten. Die Vorteile liegen in dem gemeinsamen Nutzen von Hardware und Lizenzen, einmaligem Aufwand und kurzen Wegen.

Es gilt Neutralität des DWH-Systems gegenüber den Vorsystemen, vor allem aber gegenüber den nachfolgenden

Business-Intelligence-Tools zu wahren. Das bedeutet, dass nur die Datenmodell-Formen genutzt werden, die für die Aufgabenstellung in den jeweiligen DWH-Layern am passendsten sind:

- nahezu drei NF im Enterprise-Layer
- kompakte multidimensionale Strukturen (Star-Schema) im End-User-Layer

DWH-Daten sind langlebig (zehn Jahre und länger), Vorsysteme ändern sich meist viel häufiger und BI-Tools werden ebenfalls öfter ausgetauscht oder es befinden sich sogar unterschiedliche BI-Tools gleichzeitig im Einsatz. Wer diese Regel nicht berücksichtigt, transportiert alle Änderungen in den Vorsystemen und bei den BI-Tools automatisch auch in das DWH.

Durchgängigkeit des Schichtenmodells für Benutzerzugriffe

Um Informationen nicht unnötig zwischen den Schichten zu kopieren und zu verdoppeln, sollten Endbenutzern Lesezugriffe bis in den Enterprise-Information-Layer hinein erlaubt sein. Hier sind neben Stammdaten auch viele interessante Referenzdaten abgelegt. Das macht das DWH zu einem echten Information-Repository, wodurch es für die Endanwender noch wertvoller wird. Meist haben die DWH-Administratoren ihre Hände auf dieser zentralen Schicht und sie argumentieren mit Security und Performance. Mit dieser restriktiven Haltung schmälern sie jedoch nur den Nutzen des DWH für die Anwender. Die vorgebrachten, meist technischen Punkte lassen sich jedoch mit anderen Mitteln lösen. Es gilt der Grundsatz: Eine Kopie der Daten im End-User-Layer ist nur sinnvoll, wenn Daten in eine kompaktere Modellform (etwa Star-Schema) oder in eine für Endbenutzer leichter verstehbare Darstellung überführt werden.

Große Bewegungsdaten-Bestände (Fakten-Daten) sollte es nur einmal im gesamten System geben und sie sollten nicht vom Enterprise-Layer in den User-View-Layer (Data Marts) „1:1" kopiert werden. Das bedeutet, dass sie im Enterprise-Layer angesiedelt sind, um sie über den End-User-Layer zu referenzieren.

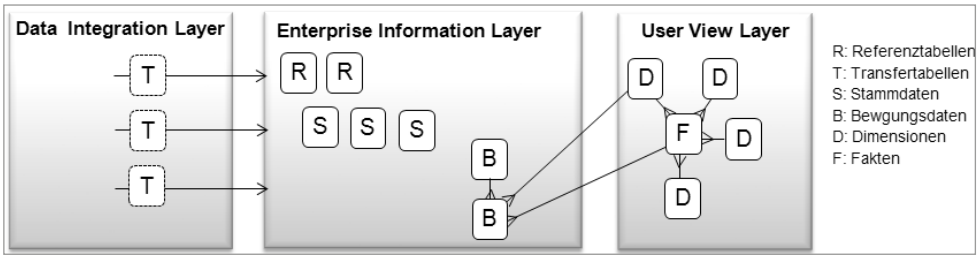


Abbildung 2: Umgang mit großen redundanten Tabellen

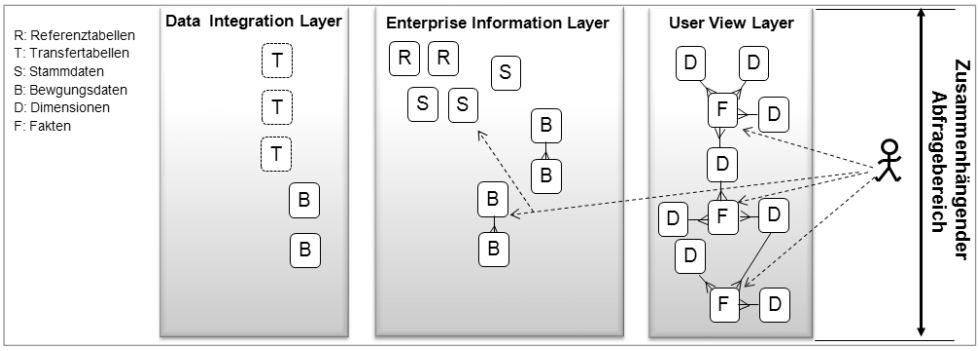


Abbildung 3: Verbundende Kennzahlen

Eventuell können die Bestände sogar mehrfach wiederverwendet werden. In Star-Schemen referenzieren bei dieser Vorgehensweise die Dimensions-Tabellen im User-View-Layer (Data Mart) als Parent Table ihre Fakten-Tabellen im Enterprise-Layer. Ohne technische Einschränkungen hinnehmen zu müssen (wie das Star-Transformation-Feature), funktioniert das jedoch nur, wenn alle Objekte in derselben Datenbank liegen (siehe Abbildung 2).

Verbundene Kennzahlen

Im User-View-Layer sind Verknüpfungen in der multidimensionalen Struktur zu schaffen, um Sachverhalte aus unterschiedlichen Kontexten und Sichten zusammenhängend abfragen zu können (siehe Abbildung 3). Die verbindenden Elemente sind die Dimensionen (Geschäftsobjekte), über die die Kennzahlen (Fakten, Measures) in einen Bezug gebracht werden können. Das verhindert sogenannte „Äpfel/Birnen-Vergleiche" schon beim Design des Datenmodells, macht sachübergreifende Abfragen gefahrlos möglich und schafft zusätzliche Flexibilität für die Anwender. Extrem formuliert: Es sollte auf alle Fakten-Tabellen mit passenden Joins über ihre Dimensionen in einer Abfrage zugegriffen werden können. Das ist in der Praxis nicht immer machbar, weil Sachgebiete unter Umständen zu unterschiedlich sind.

Vorgedachte Aggregate und Kennzahlen

Die meisten Kennzahlen sind bereits bekannt. Man bereitet sie schon in der DWH-Datenbank als vorgedachte Aggregate vor und nicht erst im BI-Tool. Das mindert den Aufwand in den aufsetzenden BI-Tools und schafft zusätzliche Flexibilität beziehungsweise mehr Angebote für die Anwender. Hier helfen Kennzahlen-Bäume, die technisch als Nested Materialized Views (aufeinander aufbauende Materialized Views) realisiert sind. Die Verwaltung erfolgt durch die Datenbank transparent, völlig automatisiert und ressourcenschonend. Der Wiederverwendungseffekt ist beachtlich. Die klassischen Aggregationstabellen sind out. Man realisiert sie heute grundsätzlich

mit Materialized Views. Das spart Verwaltungs- und ETL-Aufwand.

Generell wird man den Feinheitsgrad der Daten in User- und Enterprise-Layer so granular wie möglich halten, denn aggregieren kann man später immer noch, aber etwas Aggregiertes lässt sich im Nachhinein nicht mehr granularer gestalten. Granulare Daten bieten mehr Auswerte-Optionen und Flexibilität für die Anwender. Datenbanken und Hardware sollten so ausgelegt sein, dass auch große Datenbestände „on the Fly" aggregiert werden können. Die Technologie hierzu ist längst vorhanden.

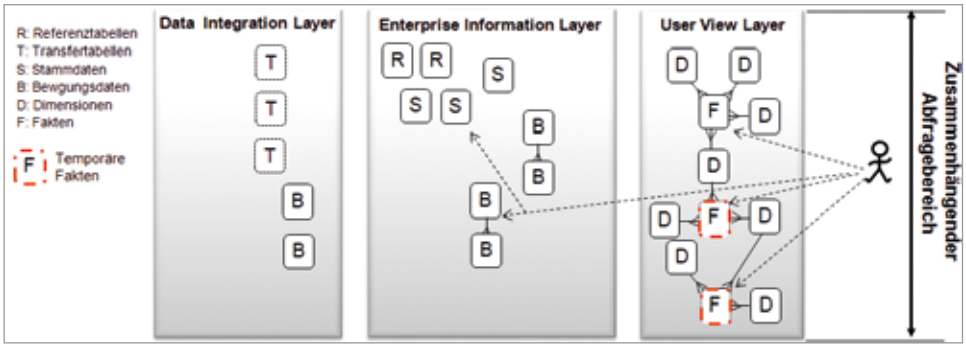


Abbildung 4: Dynamisch erstellte Auswerte-Strukturen

Dynamische Auswerte-Strukturen

Die Strukturen im User View Layer (Data Marts) sollten so gewählt werden, dass man ihre Inhalte jederzeit dynamisch aus den Daten der Vorschicht wieder herleiten kann, wenn diese durch die Anwender gezielt angefordert werden (siehe Abbildung 4). So muss man nicht permanent alle User-View-Daten bereithalten oder aktualisieren. Das spart vor allem Plattenplatz sowie ETL-Aufwand und auch ein Backup der Data-Mart-Daten ist nicht mehr nötig.

Das Security-Konzept des Analyse-Systems sollte in der DWH-Datenbank gelöst sein und nicht in dem nachgelagerten BI-Tool. Damit ist eine Offenheit für flexible BI-Strategien gegeben. Unterschiedliche BI-Tools, aber auch beispielsweise Excel- oder sonstige OCI-Anwendungen partizipieren von dem „Einmal-Security"-Aufwand in der Datenbank. Außerdem vermeidet man peinliche Security-Pannen. Security- oder Mandanten-Anforderungen dürfen nicht zu einer Verdoppelung von Strukturen und Daten führen.

PROLICENSE®
OPTIMIZING SOFTWARE ASSETS
Kompetent – Unabhängig – Erfolgsbasiert

SO RICHTIG ÜBERLIZENSIERT?
Sprechen Sie mit uns!

Wir sind nur unseren Mandanten verpflichtet.
> Compliance sichern
> Audit vermeiden
> Kosten senken

ProLicense GmbH
Friedrichstraße 191 | 10117 Berlin
Tel: +49 (0)30 60 98 19 230 | www.prolicense.com

ETL-Prozesse sollten innerhalb der Datenbank stattfinden, um unnötigen Datentransport zwischen Datenbank- und ETL-Server zu vermeiden. Zudem nutzt man die Hardware-Ressource der Datenbank sowie die Datenbank-Lizenzen besser aus. Hinzu kommen die vielen technischen Mittel der Datenbank wie Table Functions, Partitioning und die Load-Features, die extrem schnelles Laden möglich machen. „1:1"-Transport-Vorgänge, also Datenbewegungen ohne Mehrwert-liefernde Transformationen, sind zu vermeiden. Alle ETL-Prozesse sollte man Set-Based, also mengenbasiert und nicht satzweise durchführen. Constraints und Indizes sind auszuschalten. Syntaktische Datenqualität wird nicht mit Datenbank-Constraints, sondern aufgrund der besseren Performance mit SQL-Mitteln gelöst.

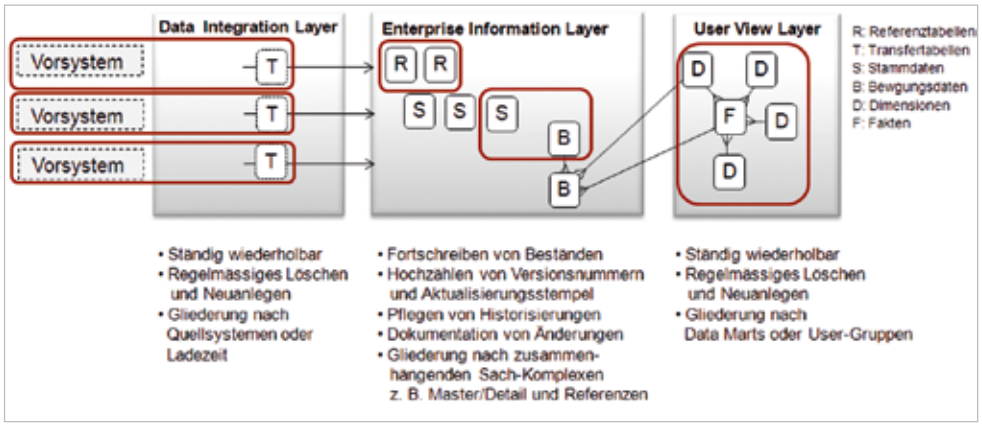


Abbildung 5: Strukturierung des ETL-Prozesses in Wiederholer- und Zusammenhangs-Gruppen

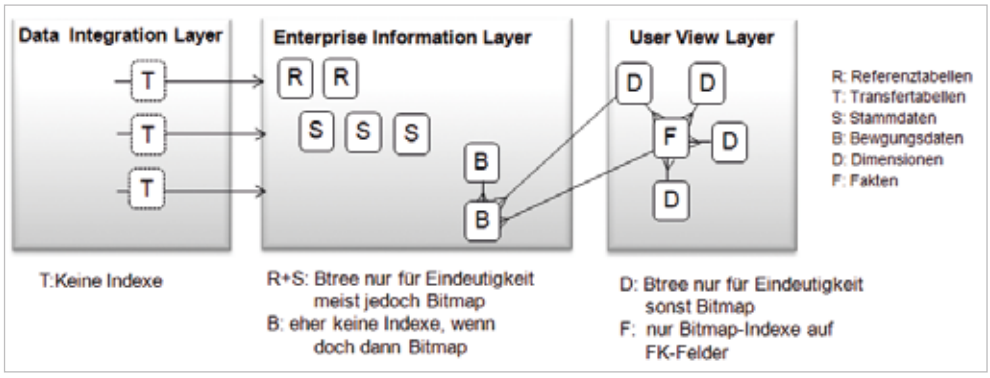


Abbildung 6: Indizierung im DWH

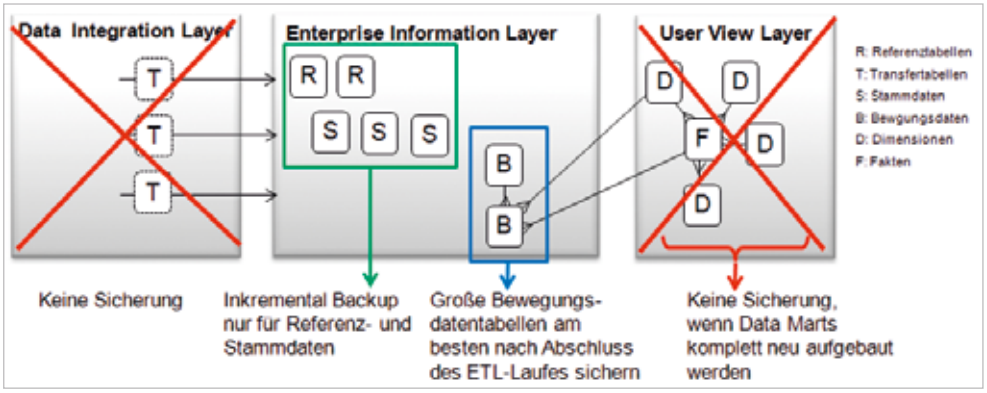


Abbildung 7: Was wird wie gesichert

Die Parallelisierung (parallele Datenbank-Prozesse) wird im ETL-Prozess nicht pauschal, sondern gezielt für einzelne Objekte gesteuert. ETL-Prozesse sind kontrollierte Vorgänge (siehe Abbildung 5). Hier weiß man in der Regel, was passiert, und kann bezogen auf Parallelisierung, aber auch bezüglich Re-Indizierung, Verwenden von Partitioning oder Aktualisierung der Statistiken kontrolliert vorgehen. Künstliches, also manuell programmiertes Parallelisieren außerhalb der Datenbank ist zu aufwändig, führt zu Lock-Situationen und sollte vermieden werden.

Kommen dennoch Engine-gestützte ETL-Werkzeuge zur Anwendung, die Transformations-Prozesse nicht in der Datenbank durchführen, sollte man die Datenleitung zwischen ETL- und Datenbank-Server möglichst leistungsstark wählen (10 GB). Einfache Datenkopien sollte man nicht mit dem ETL-Werkzeug, sondern mit Datenbank-Mitteln lösen und stattdessen einen grafischen Platzhalter für diese Datenbank-Operation in das grafische Modell einfügen.

Dokumentation, Metadaten und Kontrolle

Kaum ein anderer Bereich wird im DWH stärker vernachlässigt, als der der Metadaten. Das liegt wahrscheinlich daran, dass die Verantwortlichen hier keinen unmittelbaren Nutzen sehen. Aber die Folgen fehlender Metadaten-Dokumentation entstehen einige Zeit später umso drastischer, wenn Chaos Einzug hält: Sachverhalte sind mehrfach und auch noch unterschiedlich gespeichert. Die meisten Nutzer (und Administratoren) wissen nur zum Teil, was gespeichert ist. Große Datenmengen liegen nutzlos brach.

Ein DWH braucht neben einer sauberen Datenmodellierung auch ein neutrales Metadaten-Repository, in dem neben den technischen Objekten der Datenbank (Tabellen, Spalten, Views,

Mappings) auch fachliche Beschreibungen der Inhalte, eine Dokumentation über Herkunft (Transformationen) und Verwendung (Nutzungsstatistik) der Daten und eine Dokumentation über Benutzer und ihre Erwartungen vorhanden sind. Es muss jederzeit bekannt sein, wer welche Daten benutzt und wo welche Daten wie gespeichert sind. Diese Information sollte für alle Beteiligten beispielsweise in einer Web-Oberfläche zur Verfügung stehen, und nicht nur dem Administrator.

Indizierung im DWH

Im DWH nutzt man eher Bitmap-Indizierung als Btrees (siehe Abbildung 6). Bitmaps benötigen weniger Platz und sind bei mengenorientierten Abfragen schneller. Btrees nutzt man oft nur zur Verwaltung der Eindeutigkeit bei Stamm-, Referenz- und Dimensions-Tabellen (Primary-Key-Funktion).

Daten-Backup des Data Warehouse

DWH-Systeme benötigen ein eigenes, also von OLTP-Systemen unabhängiges Backup-Verfahren, denn nicht alles muss man regelmäßig sichern. Integration und User-View-Layer sind nicht zu sichern. Referenz- und Stammdaten muss man nur dann sichern, wenn sie sich geändert haben (Inkrementelles Backup des RMAN). Große Bewegungsdaten-Tabellen sichert man gekoppelt an den ETL-Prozess, weil man hier die geänderten Datenbestände kennt, etwa in Verbindung mit Partitioning (siehe Abbildung 7). Backups nur über das Storage-System (Image-Copy) sind zu vermeiden (Gefahr von Block-Corruption und zu teuer).

Regeln aus Hardware-Sicht

Data-Warehouse-Daten sollten in einem separaten DWH-Storage-System liegen. Eine gemeinsame Storage-Nutzung, zum Beispiel mit OLTP-Systemen, ist zu vermeiden. Das DWH liest meist größere und zusammenhängende Datenmengen, oft sogar über einen Full Table Scan, OLTP-Systeme dagegen eher in kleineren Daten-Häppchen und dazu oft noch über einen Index. Das gilt analog auch für das Netzwerk zwischen DWH-Storage und Datenbank-Server. Direkt angeschlossener Datenspeicher ist sicher das beste (Exadata – Appliance-Prinzip).

Bei der Wahl des Storage-Systems setzt man eher auf mehr und dafür kleine Platten. Ideal ist es, wenn diese nur zur Hälfte beschrieben werden (Datenablage nur auf den äußeren Spindeln). Der Hauptspeicher sollte mindestens mit 8 GB pro Core dimensioniert sein, um aufwändige Star-Join-Operationen und analytische Funktionen besser zu unterstützen.

Die Menge der CPU-Cores richtet sich nach der zu lesenden Datenmenge pro Sekunde (MB/sec). Man geht bei heutigen CPUs (>2GHz) von etwa 200 MB/sec Verarbeitungskapazität pro Core aus. Bei einem optimierten 30TB-DWH mit etwa 5000 MB/sec Datendurchsatz (5GB/sec) sind das ungefähr 25 Cores und etwa 100 Platten (ohne Spiegelung). Wer weniger als 5 GB/sec Lese-Performance in einem solchen DWH hat, sollte sich damit nicht zufrieden geben und Alternativen suchen.

Die Kosten für Storage-Systeme variieren sehr stark. Hat man privaten Storage für das DWH gewählt, so kann man unterschiedlich teure Platten für unterschiedlich wichtige/aktuelle Daten im DWH einsetzen. Meist wird nur ein geringer Teil der DWH-Daten intensiv genutzt und der große Rest liegt für Monate brach.

Alfred Schlaucher
alfred.schlaucher@oracle.com

Automatisierte SAP-Systemkopien auf Knopfdruck

Libelle SystemCopy

- Ohne in Ihre SAP-Umgebung einzugreifen bzw. diese zu verändern
- Ohne aufwändige Vorplanung
- Mit minimaler Durchlaufzeit
- Bei gleichbleibender Qualität der Kopie

... mit deutlich reduzierten Prozesskosten

Hans-Joachim Krüger
Chief Technology Officer
Libelle AG

Erfahren Sie mehr:
www.Libelle.com/systemcopy

ORACLE Gold Partner

Libelle

Libelle AG
Gewerbestr. 42 • 70565 Stuttgart, Germany
T +49 711 / 78335-0 • F +49 711 / 78335-148
www.Libelle.com • sales@libelle.com

Mitten in der Nacht bricht die ETL-Verarbeitung ab, weil ein falscher oder unvollständiger Datensatz geliefert wurde. Das Laden des Data Warehouse ist nicht möglich, und erst nach einem manuellen Eingriff können die Ladeprozesse fortgesetzt oder neu gestartet werden. Das muss nicht sein! Zwischen ETL-Abbruch und Ignorieren der fehlerhaften Datensätze gibt es zahlreiche weitere Möglichkeiten, um häufig auftretende Fehlersituationen zu erkennen und automatisch zu behandeln.

Fehlertolerante Ladeprozesse in Oracle gegen schlaflose Nächte

Dani Schnider, Trivadis AG

Oft sind es Kleinigkeiten, die zum Abbruch der ETL-Verarbeitung führen. Ein fehlendes Attribut, ein unbekannter Codewert, eine ungültige Referenz auf eine Dimension oder eine Schlüsselverletzung aufgrund doppelt oder mehrfach gelieferter Datensätze kann dazu führen, dass der ETL-Job abgebrochen wird. Die Folge davon ist entweder, dass die aktuellen Daten am anderen Morgen nicht zur Verfügung stehen oder dass ein Mitarbeiter des Betriebsteams beziehungsweise ein DWH-Entwickler in der Nacht manuelle Daten-Korrekturen vornehmen und den ETL-Job wieder starten muss. Im ersten Fall führt dies dazu, dass bei wiederkehrenden Fehlerfällen die Akzeptanz des Data Warehouse bei den Anwendern stark leidet. Der zweite Fall führt zu häufigen manuellen Eingriffen in der Nacht und somit zu schlaflosen Nächten bei den betroffenen DWH-Mitarbeitern.

```
INSERT INTO cls_products (pro-
duct_code, product_desc)
SELECT product_code, product_
desc
  FROM stg_products;

ORA-01400: cannot insert
NULL into("CLEANSE"."CLS_
PRODUCTS"."PRODUCT_DESC")
```

Listing 1

Weder unzufriedene Benutzer noch übermüdete DWH-Entwickler sind dem Erfolg eines Data Warehouse förderlich. Deshalb brauchen wir Alternativen, mit denen die ETL-Verarbeitung auch dann fortgesetzt werden kann, wenn einzelne Datensätze unvollständig oder falsch sind. Ein einfacher Ansatz ist, Fehler einfach zu ignorieren. So ist es beispielsweise in Oracle Warehouse Builder möglich, die Konfiguration eines Mappings so einzustellen, dass eine maximale Anzahl von Fehlern pro Lauf erlaubt ist. Das klingt zwar verlockend einfach, hat aber zwei Nachteile. Erstens werden die Daten unvollständig geladen. Zweitens dauert die Verarbeitung bei großen Daten-Mengen länger, weil das Mapping in der Betriebsart „row-based" ausgeführt werden muss. In vielen DWH-Systemen mag dieser Ansatz genügen, aber je nach Anforderungen an Daten-Qualität und Ladezeiten sind solche Lösungen nicht realistisch. Zum Glück gibt es andere, ebenfalls einfache Lösungen.

Auf den folgenden Seiten wird anhand von typischen Fehler-Situationen aufgezeigt, wie fehlerhafte Datensätze erkannt und so behandelt werden können, dass die ETL-Verarbeitung trotzdem fortgesetzt werden kann. Die hier vorgestellten Verfahren lassen sich in Oracle sowohl mit SQL als auch mit ETL-Tools wie Oracle Warehouse Builder (OWB) oder Oracle Data Integrator (ODI) realisieren. In den nachfolgenden Beispielen wird jeweils der Lösungsansatz mit SQL gezeigt, der aber sinngemäß auch mit OWB oder ODI implementiert werden kann.

Fehlende Attribute

Ein typischer Fehlerfall ist ein leeres Attribut, das in der Zieltabelle als „NOT NULL" definiert ist. Dies führt normalerweise zu einem Abbruch der Verarbeitung, sobald ein unvollständiger Datensatz, beispielsweise ein Produkt ohne Beschreibung, auftritt. Das folgende einfache Beispiel zeigt, dass die Produkt-Dimension aus der Stage-Tabelle „STG_PRODUCTS" in die bereinigte Tabelle „CLS_PRODUCTS" der Cleansing-Area lädt. Ohne explizite Fehlerbehandlung in SQL führt dies im Falle eines NULL-Werts zu einem Abbruch mit entsprechender Fehlermeldung (siehe Listing 1).

Die einfachste Lösung, um einen Abbruch zu verhindern, besteht darin, die fehlerhaften Datensätze zu filtern und somit zu vermeiden, dass sie in die Ziel-Tabelle geschrieben werden. Dies kann entweder mit einem Cursor-Loop und entsprechendem Exception Handler in PL/SQL implementiert werden (also row-based) oder ganz einfach mit einer WHERE-Bedingung im entsprechenden SQL-Statement (siehe Listing 2).

Bei dieser Lösung nehmen wir in Kauf, dass unter Umständen unvollständige Daten ins Data Warehouse geladen werden. In einigen Fällen kann

dies akzeptabel sein, solange die Anzahl der Fehler nicht zu groß wird. Es empfiehlt sich deshalb, nach dem Laden in die Cleansing-Area einen Check einzubauen, der die Anzahl der Datensätze in Staging- und Cleansing-Area vergleicht und bei Überschreitung eines Schwellwerts die Verarbeitung abbricht. Diese Anforderung lässt sich – sofern der Schwellwert als absolute Zahl und nicht als prozentualer Anteil definiert werden soll – in Oracle mithilfe von DML-Error-Logging implementieren (siehe Listing 3).

Dabei werden Datensätze, die zu Constraint-Verletzungen führen, nicht in die Ziel-Tabelle geladen, sondern in eine zuvor erstellte Fehler-Tabelle geschrieben. Bei Bedarf kann ein Schwellwert – hier maximal 100 Fehler – definiert werden. Ein wesentlicher Vorteil von DML-Error-Logging besteht darin, dass die Verarbeitung weiterhin „set-based“ ausgeführt werden kann.

Da aber die fehlerhaften Datensätze nicht in die Ziel-Tabelle geladen werden, handelt es sich hier nur um eine erweiterte Implementation der Filter-Variante – allerdings mit dem Vorteil, dass die fehlerhaften Datensätze nicht einfach ignoriert werden, sondern in der Fehler-Tabelle verfügbar sind.

Das Filtern von fehlerhaften Datensätzen – ob mit „WHERE“-Bedingung oder DML-Error-Logging implementiert – führt spätestens dann zu Problemen, wenn Referenzen auf die fehlenden Daten auftreten. Was passiert in unserem Beispiel, wenn in einem späteren Schritt der ETL-Verarbeitung Fakten in eine Fakten-Tabelle geladen werden, die auf das fehlende Produkt verweisen? Das Filtern des fehlerhaften Datensatzes kann zu Folgefehlern in weiteren ETL-Schritten führen.

Ein anderer einfacher, aber nicht zu empfehlender Ansatz ist, den „NOT NULL“-Constraint auf der Ziel-Tabelle wegzulassen und somit fehlende Attributwerte zu erlauben. Das hört sich zwar verlockend simpel an, führt aber ebenfalls zu Folgefehlern und zu Problemen bei den Auswertungen. Wie sollen die leeren Felder in einem Report oder einem OLAP-Tool angezeigt werden, sodass sie für den Anwender erkennbar sind? Zusätzliche Fallunterscheidungen in den Abfragen werden notwendig, um die leeren Felder entsprechend zu markieren. Damit wir dies nicht für jede Auswertung tun müssen, ist es naheliegender und einfacher, diesen Schritt bereits im ETL-Prozess durchzuführen. Dies führt zum empfohlenen und bewährten Lösungsansatz, mit Singletons zu arbeiten.

Ein Singleton ist ein Platzhalter oder Defaultwert, der in bestimmten Fehlersituationen, beispielsweise bei einem leeren Attribut, eingesetzt wird. Für unser Beispiel möchten wir anstelle einer leeren Produktbezeichnung den Wert „unknown“ anzeigen. Das lässt sich durch eine einfache Erweiterung im ETL-Prozess erreichen (siehe Listing 4).

```
INSERT INTO cls_products (pro-
duct_code, product_desc)
SELECT product_code, product_
desc
  FROM stg_products
 WHERE product_desc IS NOT
NULL;
```

Listing 2

```
INSERT INTO cls_products (pro-
duct_code, product_desc)
SELECT product_code, product_
desc
  FROM stg_products
LOG ERRORS REJECT LIMIT 100;
```

Listing 3

```
INSERT INTO cls_products (product_code, product_desc)
SELECT product_code
     , NVL(product_desc, 'Unknown')
  FROM stg_products;
```

Listing 4

Dieses Verfahren, das wir in erweiterter Form auch für Code-Lookups und Referenzen auf Dimensionen verwenden können, wird in vielen Data Warehouses eingesetzt und hat den Vorteil, dass der Datensatz geladen und somit später referenziert werden kann, beispielsweise von Fakt-Einträgen mit Verweisen auf den Dimensions-Eintrag. In Reports und OLAP-Tools werden die Einträge auch angezeigt, allerdings mit der Bezeichnung „unknown“ statt der korrekten (fehlenden) Bezeichnung.

Wie verhalten sich Singletons in Kombination mit Slowly Changing Dimensions (SCD)? Nehmen wir an, das fehlende Attribut wird im Quellsystem nachträglich ergänzt und in einem späteren ETL-Lauf ins DWH geladen. Bei „SCD Typ 1“ wird der bisherige Datensatz überschrieben, und ab diesem Zeitpunkt wird in allen Auswertungen der korrekte Wert angezeigt. Bei „SCD Typ 2“ wird eine neue Version in die Dimensionstabelle eingefügt. Das hat zur Folge, dass neue Fakten auf den vollständigen Eintrag verweisen. Bereits geladene Fakten zeigen aber weiterhin auf den Eintrag mit der Bezeichnung „unknown“.

Unbekannte Code-Werte

Typisch in ETL-Prozessen sind Lookups auf Code- oder Referenz-Tabellen. Anhand eines Code-Werts oder eines fachlichen Schlüssels werden ein künstlicher Schlüssel (Surrogate Key) sowie eventuell weitere Attribute wie eine Bezeichnung ermittelt. Was passiert nun während der ETL-Verarbeitung, wenn der entsprechende Code-Wert in der Lookup-Tabelle nicht vorhanden ist? Der einfachste Fall besteht auch hier wieder darin, die fehlerhaften Datensätze zu ignorieren.

Dieses Verfahren wird häufig – oft ungewollt – verwendet, indem in SQL ein normaler Inner-Join zwischen Quell-Tabelle und Lookup-Tabelle gemacht wird. Die Folge ist, dass Datensätze mit fehlenden oder unbekannten

```
INSERT INTO cls_products (product_code, product_desc, category_id)
SELECT stg.product_code
     , NVL(product_desc, ‚Unknown‘)
     , NVL(lkp.category_id, -1)
FROM stg_products stg
  LEFT OUTER JOIN co_categories lkp
    ON (lkp.category_code = stg.category_code);
```

Listing 5

```
INSERT INTO cls_products (product_code, product_desc)
SELECT product_code, NVL(product_desc, ‚Unknown‘)
  FROM (SELECT product_code, product_desc
             , ROW_NUMBER() OVER(PARTITION BY product_code
                                 ORDER BY product_desc) rnum
          FROM stg_products)
 WHERE rnum = 1;
```

Listing 6

ID	CODE	DESCRIPTION
-1	n/a	Unknown

Tabelle 1

ID	CODE	DESCRIPTION
5432	ABC	Unknown

Tabelle 2

Code-Werten nicht in die Ziel-Tabelle geschrieben werden. Weil durch diese Lösungsvariante fehlerhafte Datensätze gefiltert werden, haben wir wiederum das Problem, dass die Daten unvollständig geladen werden. Um dies zu vermeiden, können auch hier Singletons eingesetzt werden, wenn auch in etwas erweiterter Form.

Beim initialen Laden des Data Warehouse wird in jede Lookup-Tabelle ein Singleton-Eintrag geschrieben, der durch einen speziellen Schlüssel, zum Beispiel eine negative ID, gekennzeichnet wird (siehe Tabelle 1).

Beim Lookup wird ein Outer-Join auf die Lookup-Tabelle gemacht. Damit ist gewährleistet, dass auch Datensätze mit fehlenden oder unbekannten Code-Werten in die Ziel-Tabelle geschrieben werden. Um zu vermeiden, dass ein leerer Schlüssel übergeben wird, wird der leere Eintrag durch den Schlüssel des Singleton-Eintrags, in unserem Fall durch den Wert „-1“, ersetzt. In Oracle SQL lässt sich dies einfach mittels der NVL-Funktion realisieren (siehe Listing 5).

Das Prinzip der Singletons erlaubt auch hier ein vollständiges Laden aller Daten, hat aber den Nachteil, dass eine nachträgliche Zuordnung zum korrekten Code-Wert nicht mehr möglich ist. Auch wenn später der fehlende Code nachgeliefert wird, kann er in den bereits geladenen Daten nicht mehr aktualisiert werden – es sei denn, der Original-Wert des Quellsystems wird zusätzlich im DWH gespeichert.

Eine flexiblere, aber auch etwas aufwändigere Möglichkeit besteht darin, fehlende Code-Werte vorgängig in die Lookup-Tabelle zu laden. Angenommen, das Quellsystem liefert einen Datensatz mit dem Code „ABC“, der im Data Warehouse noch nicht vorhanden ist. Deshalb wird nun ein neuer Eintrag in die Code-Tabelle geschrieben, der im ersten Moment aussieht wie ein Singleton-Wert, jedoch einen neuen Surrogate Key (hier den Wert „5432“) zugewiesen bekommt (siehe Tabelle 2).

Wird die Bezeichnung für den Code „ABC“ später nachgeliefert, kann der Datensatz überschrieben werden. Da die bereits geladenen Daten auf die ID „5432“ verweisen, wird ab diesem Zeitpunkt der korrekte Wert angezeigt.

Die zuvor eingefügten Datensätze in der Lookup-Tabelle entsprechen somit den echten Datensätzen, die später vom Quellsystem geliefert werden. Da sie bereits vorhanden sind, aber noch keinen Namen (also keine Bezeichnung) haben, nennen wir sie „Embryo“-Einträge. Die Geburt, das heißt die Umwandlung eines Embryo-Eintrags in einen echten Lookup-Eintrag, erfolgt, wenn der Code-Wert vom Quellsystem geliefert und ins Data Warehouse geladen wird.

Fehlende Dimensions-Einträge

Was hier anhand von Code-Tabellen beschrieben wurde, lässt sich auf beliebige Lookup-Tabellen und somit auch auf Dimensions-Tabellen anwenden. Dies ist dann interessant, wenn die Situation auftreten kann, dass Fakten bereits geliefert werden, bevor die zugehörigen Dimensions-Einträge im Data Warehouse vorhanden sind. Auch hier können die oben beschriebenen Varianten-Filterung, Singleton- und Embryo-Einträge angewendet werden. Eine ausführliche Beschreibung dieser Problemstellung sowie der drei Lösungsvarianten ist im Artikel „Wenn die Fakten zu früh eintreffen“ (siehe http://www.trivadis.com/uploads/tx_cabagdownloadarea/Wenn_die_Fakten_zu_frueh_eintreffen.pdf) dokumentiert.

Doppelte Datensätze

Eine ebenfalls häufig anzutreffende Fehlerursache sind doppelt oder mehrfach vorhandene Datensätze innerhalb einer Lieferung. Grund dafür können Mehrfach-Lieferungen oder nicht eindeutige Join-Kriterien in den Extraktionsprozessen sein. Doppelte Datensätze führen typischerweise zu Schlüsselverletzungen (Primary Key Violation oder Unique Key Violation) beim Laden ins DWH und somit zu einem Abbruch der ETL-Verarbeitung.

Der nächstliegende und einfachste Ansatz, um doppelte Datensätze zu eliminieren, besteht darin, ein „DISTINCT“ in der Abfrage auf die Stage-Tabelle zu verwenden. Solange alle Attribute der Datensätze identisch sind, funktioniert dies tadellos. Doch sobald sich die Datensätze in mindestens ei-

Unsere Inserenten

ARETO Consulting GmbH www.areto-consulting.de	S. 7
Apps Associates www.appsassociates.de	S. 41
Heise-Verlag www.heise.de	U 3
Hunkler GmbH & Co. KG www.hunkler.de	S. 3
KeepTool GmbH www.keeptool.com	S. 21
Libellle AG www.libelle.com	S. 15
MT-AG www.	S. 33
MuniQsoft GmbH www.muniqsoft.de	S. 45
OPITZ CONSULTING GmbH www.opitz-consulting.com	U 2
ProLicense GmbH www.prolicense.com	S. 13
Trivadis GmbH www.trivadis.com	U 4

nem beschreibenden Attribut unterscheiden, aber dennoch den gleichen Schlüssel besitzen, wird die ETL-Verarbeitung trotzdem abgebrochen. Dieser Fall dürfte zwar theoretisch nicht auftreten, kommt aber in der Praxis aufgrund von ungenauen oder fehlerhaften Extraktionsprozessen leider vor. Um sicherzustellen, dass auch in diesem Fall nur ein Datensatz übernommen wird, können beispielsweise mit der analytischen Funktion „ROW_NUMBER“ alle Datensätze mit einem Schlüssel durchnummeriert und nur der jeweils erste Datensatz übernommen werden, wie Listing 6 zeigt.

Diese Variante funktioniert in jedem Fall und durch Verwendung der analytischen Funktion erst noch effizient. Doch die entscheidende Frage lautet: Welcher ist der erste Datensatz? Im hier aufgeführten Beispiel wird der Eintrag übernommen, dessen Produktbezeichnung alphabetisch zuoberst erscheint. Die Regel ist völlig willkürlich und dient nur dazu, einen zufälligen Datensatz zu übernehmen. Falls eine solche pragmatische Lösung nicht genügen sollte, ist die Variante mit „DISTINCT“ besser geeignet, da sie zwar bei identischen Datensätzen funktioniert, aber bei unterschiedlichen Datensätzen mit gleichem Schlüssel zu einem Abbruch führt. Eine elegantere Lösung besteht darin, mithilfe von DML-Error-Logging die doppelten Datensätze in die zuvor erstellte Fehler-Tabelle zu schreiben (siehe Listing 7).

```
INSERT INTO cls_products (product_code, product_desc)
SELECT product_code, product_desc
  FROM stg_products
LOG ERRORS REJECT LIMIT UNLIMITED;
```

Listing 7

```
SELECT product_code, ora_err_mesg$ FROM err$_cls_products;

PRODUCT_CODE   ORA_ERR_MESG$
-------------- ------------------------------------------------
12345-67890-76 ORA-00001: unique constraint
               (CLEANSE.CLS_PRODUCTS_UK) violated
```

Listing 8

Nehmen wir an, in der Stage-Tabelle „STG_PRODUCTS“ kommen zwei Datensätze mit dem Produktcode „12345-67890-76“ vor. Der erste Datensatz wird dann in die Ziel-Tabelle geladen, der zweite in die Fehler-Tabelle, wie die Abfrage auf die Fehler-Tabelle zeigt (siehe Listing 8).

Welcher der erste (und damit korrekte) Datensatz ist, hängt von der physischen Speicherung in der Stage-Tabelle ab und ist somit auch zufällig. Das oben beschriebene Problem im Falle von unterschiedlichen Attributen tritt hier also genauso auf wie bei der Variante mit „ROW_NUMBER“.

Fazit

Eine allgemeine Patentlösung für die Behandlung von Fehlern in ETL-Prozessen gibt es zwar nicht und es ist auch kaum möglich, alle auftretenden Fehlerfälle abzufangen und automatisch zu behandeln. Trotzdem sollte versucht werden, durch geeignete Verfahren die häufig auftretenden Fehlerfälle so zu behandeln, dass sie nicht zu einem Abbruch der ETL-Verarbeitung führen.

Bei allen hier beschriebenen Verfahren kann die Verarbeitung auch beim Auftreten von fehlerhaften Datensätzen fortgesetzt werden, und alle Lösungen lassen sich set-based ausführen. Jede Variante hat aber Vor- und Nachteile bezüglich Daten-Qualität, Folgefehlern und Komplexität. Deshalb sollten für jedes Data Warehouse die geeigneten Verfahren zur Fehlerbehandlung definiert werden. Die in diesem Artikel beschriebenen Methoden können dabei helfen, stabilere ETL-Abläufe zu implementieren und somit den Betriebsverantwortlichen und DWH-Entwicklern eine ruhige und erholsame Nacht zu ermöglichen.

Dani Schnider
dani.schnider@trivadis.com

Wie oft in unserem Leben wünschen wir uns, bestimmte wiederkehrende Routineaufgaben nicht mehr von Hand oder am liebsten gar nicht mehr selbst erledigen zu müssen? Warum soll es bei der Entwicklung von ETL-Prozessen in einem Data Warehouse anders sein? Nach einer Reihe durchgeführter Projekte haben wir uns diese Frage gestellt und ein mit Oracle- und OWB-Mitteln entwickeltes Framework erstellt, das es ermöglicht, Mappings im Oracle Warehouse Builder automatisch zu generieren.

Automatische Generierung von OWB Mappings: mehr Zeit für das Wesentliche

Irina Gotlibovych, MT AG

Bei der Entwicklung der ETL-Prozesse in einem Data Warehouse sieht man sich wiederholt vor die Aufgabe gestellt, Prozess-Schritte aufbauen zu müssen, die einer gleichartigen Logik folgen. So werden in jedem Projekt viele Daten-Objekte auf die gleiche Weise aus Quellsystemen in den Arbeitsbereich geladen. Der Transformations-Schritt führt Daten in das einheitliche Format der Zieldatenbank über; gängige Verfahren dabei sind beispielsweise Datentyp-Konvertierung und Daten-Bereinigung. Anschließend werden Daten nach dem gleichen Prinzip – wie etwa Delta Load oder SCD – in das Data Warehouse eingebracht. In der Praxis bedeutet dies oft, dass logisch identische Mappings in manueller Kleinarbeit angelegt werden. In jedem dieser Mappings sind von Hand Operatoren anzulegen und zu verbinden. Für jedes Attribut eines Expression Operators muss man manuell den Ausdruck eintragen. Eigenschaften von Operatoren und Attributen sind immer wieder neu zu setzen. Kommt in einer Quelltabelle später eine neue Spalte hinzu, muss sie in den meisten Fällen identisch zu den anderen Spalten geladen und verarbeitet werden. Um das zu erreichen, ist aber im entsprechenden Mapping jeder betroffene Operator manuell zu ändern. Besonders aufwändig wird es, wenn sich die grundlegende Logik ändert; im Falle von späteren Änderungsanforderungen muss zumeist jedes Mapping wieder angepasst werden. Obwohl sie der gleichen Logik folgen, ist trotzdem jedes dieser Mappings einzeln zu testen: Da sie unabhängig voneinander entwickelt wurden, können in jedem auch unterschiedliche Fehler auftreten. Bei der manuellen Entwicklung spielt der menschliche Faktor eine enorme Rolle. Sich wiederholende Entwicklungsarbeiten wie das fünfzigfache Anfertigen eines Delta-Load-Mappings sind monoton und führen dadurch zu Fehlern.

Generische ETL-Entwicklung mit dem OWB

Es stellt sich die Frage: „Warum entsteht so ein Mehraufwand bei der manuellen Entwicklung und wie kann man diesen vermeiden?“ Wäre es nicht schöner, die Logik nur einmal zu entwickeln und diese dann in weiteren Mappings beziehungsweise Projekten mehrmals zu verwenden? Das Problem bei der Entwicklung im OWB besteht darin, dass es keine Möglichkeit gibt, Mappings ohne Bindung an konkrete Objekte (Tabellen, Spalten etc.) anzulegen. Die fachliche Logik eines Mappings ist immer fest mit den Umgebungs-Informationen verbunden. Um die gleiche Logik nicht mehrfach neu erzeugen zu müssen, wäre ein Weg erforderlich, Mappings generisch, also ohne Bezug zu den eigentlichen Objekten definieren zu können. Die Erzeugung der Mappings kann dann automatisch erfolgen, wobei Objekt-Namen als Parameter dem Generierungsprozess mitgegeben werden. Dieser Ansatz liegt bei der Entwicklung des OWB Mapping Generators zugrunde.

Welche Vorteile bringt so ein generischer Ansatz? Angenommen, man möchte Slowly Changing Dimensions Typ 2 in seinem Data Warehouse implementieren. Bei der manuellen Entwicklung würde man für jede Ziel-

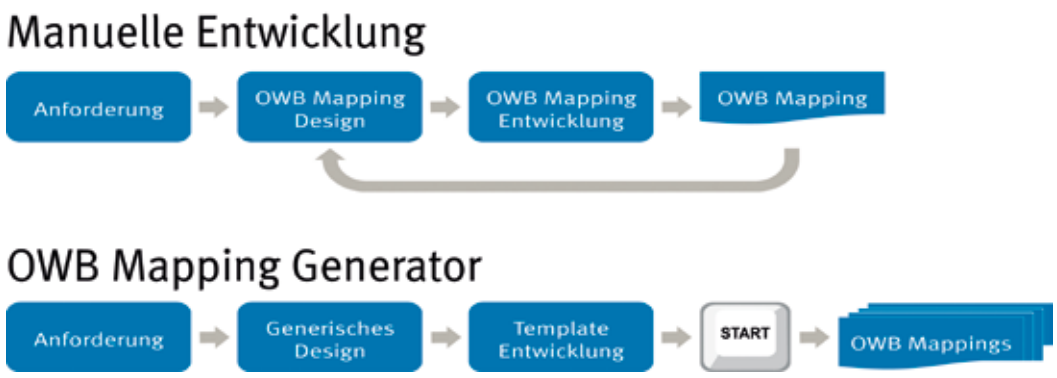


Abbildung 1: Prozesskette bei der Erstellung von Mappings mit dem OWB Mapping Generator im Vergleich zur manuellen Entwicklung

Tabelle die komplexen Join- und Splitterbedingungen in Abhängigkeit von den jeweiligen Primärschlüsseln und Spalten einzeln implementieren. Wählt man den generischen Ansatz, wird die Logik unabhängig von Tabellen und Spalten einmal in Form eines Templates definiert. Der Generierungsprozess arbeitet nun nach allgemeinen Regeln, was identischen Code und gleiche Qualität für alle Mappings garantiert. Und wenn später eine neue Spalte hinzukommt? Da das Template allgemein definiert wurde und die konkreten Primärschlüssel beziehungsweise Spalten erst bei der Verarbeitung aus der Datenbank ausgelesen werden, wird die neue Spalte bei einer erneuten Generierung des Mappings automatisch berücksichtigt. Es ist keine manuelle Anpassung des Mappings im OWB Design Center notwendig. Und wie sieht es mit Fehlerbehebung und Testen aus? Da man die Logik an einer Stelle entwickelt, müssen die Fehler auch nur an einer Stelle behoben werden – nämlich in dem zugrunde liegenden Template und nicht in jedem Mapping einzeln. Ist man einmal sicher, dass die Logik in dem Template richtig definiert ist, kann man auch sicher sein, dass jedes damit generierte Mapping korrekt laufen wird.

OWB Mapping Generator

Der OWB Mapping Generator ist ein kleines Framework, mit dem sich Mappings im Oracle Warehouse Builder auf Basis von mitgelieferten oder selbstentwickelten Templates automatisch generieren lassen. Die Standard-Implementierung basiert auf Oracle 11*g* R2 und Oracle Warehouse Builder 11*g* R2 Basic ETL, eine Anpassung für andere Versionen ist aber problemlos möglich. Bei der Entwicklung wurde Wert auf die Verwendung von Standard-OWB-Features gelegt, um zusätzliche Lizenzkosten zu vermeiden. Im Gegensatz zur manuellen Entwicklung setzt man beim Gebrauch des Frameworks nicht mehr jedes Mapping im OWB Design Center einzeln um, sondern definiert ein allgemeingültiges Template für eine „Klasse“ von Mappings (siehe Abbildung 1). Anschließend generiert man die dazugehörigen Mappings unter Einbeziehung der Projektvorgaben automatisch. An dieser Stelle ist besonders anzumerken, dass es sich bei den Templates nicht um ein programmiertes TCL-Skript zur Generierung der Mappings handelt, sondern genauso wie im OWB Design Center um eine deklarative Definition auf Basis von Metadaten (siehe Abbildung 2).

Die Generierung der Mappings wird über einen „OWB Expert“ aus der Oracle Warehouse Builder GUI angestoßen. Dieser leitet dialoggestützt durch die einzelnen Schritte. Nachdem man seine Auswahl bezüglich des zu generierenden Templates und der zu verwendenden Objekte getroffen hat, legt der OWB Mapping Generator die Eingaben in den Steuer-Tabellen auf

KeepTool mit neuer Version 10.1

Das handliche Werkzeug für Oracle™-Datenbanken

Zahlreiche neue Funktionen, z.B.

- Erzeugen von AWR-Reports bei vorhandener *Lizenz des Oracle Datenbank Diagnostic Packs.*
- Schnelle Textsuche quer über alle Tabellen im Schema.
- Praktische dynamische Tooltip-Hinweise im DataContent.
- Mehrstufige Pivot-Ansicht mit Visualisierung im DataContent.

Laden Sie die kostenlose Testversion unter www.keeptool.com herunter.

ORACLE Gold Partner

keeptool

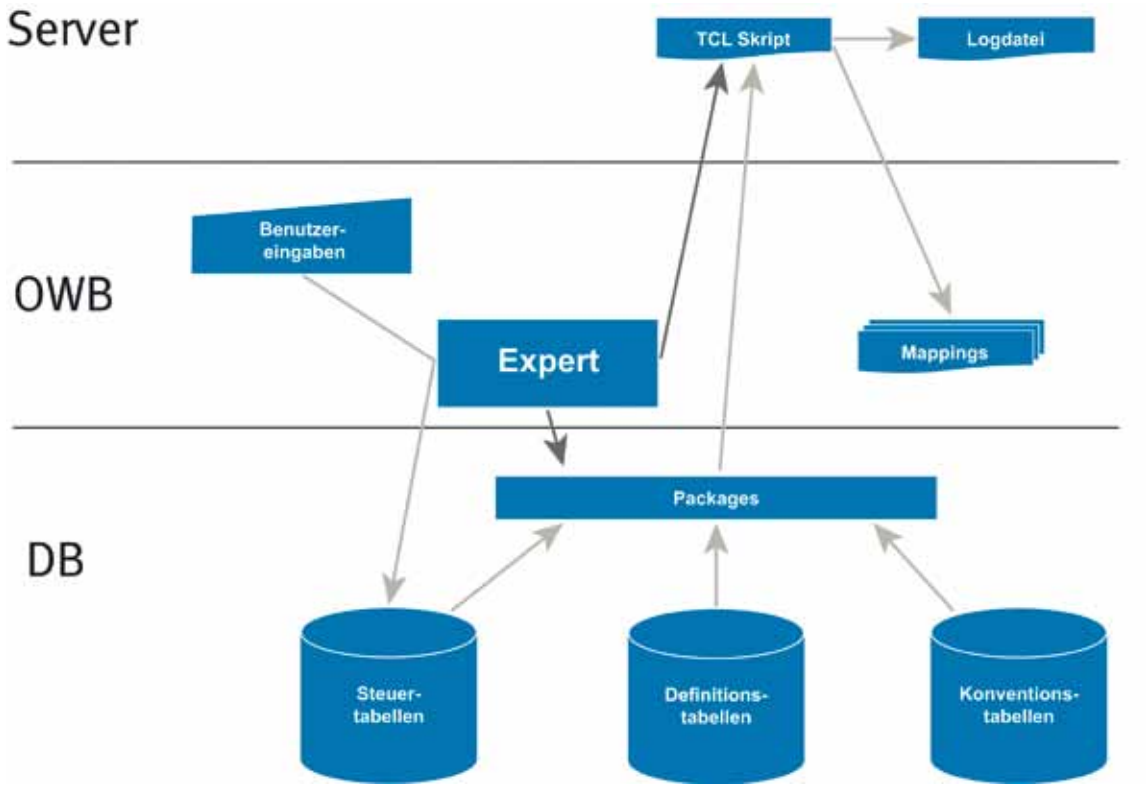


Abbildung 2: Architektur des Frameworks

der Datenbank ab. Danach werden PL/SQL-Prozesse gestartet (siehe Abbildung 2), die aus den Definitions-Tabellen das gewünschte Template auslesen und die darin gespeicherte Logik mit den Umgebungs-Informationen aus den Konventions-Tabellen anreichern. Damit wird nun ein TCL (OMB Plus) Skript generiert, mit dem die Mappings anschließend automatisch erzeugt werden. Das Ergebnis ist sofort im Oracle Warehouse Builder sichtbar und kann weiterverwendet werden. Während der Generierung der Mappings wird jeder Schritt in einer Logdatei auf dem Server protokolliert. Man hat so stets den vollen Überblick über die generierten Objekte.

Ein Template statt vieler Mappings

Kernstück der Architektur des OWB Mapping Generators bilden die bereits mehrfach erwähnten generischen Templates. Der OWB Mapping Generator stellt in der Datenbank einen Satz von Definitions-Tabellen bereit, in denen die Templates mithilfe der Metadaten beschrieben werden. In den Definitions-Tabellen findet man keine Tabellen- oder Attribut-Namen; die konkreten Objekte werden erst während der Generierung automatisch an die Templates gebunden. Die zentralen Tabellen des Datenmodells enthalten Informationen über Operatoren, Attribute, Properties und Connections. Die Begrifflichkeiten im OWB Mapping Generator sind gleich wie im Oracle Warehouse Builder – man findet sich demnach schnell zurecht. So lassen sich für einen Operator unter anderem Name, Typ und OWB-Modul angeben. Bei den Namen für Tabellen-Operatoren, die sich von Mapping zu Mapping unterscheiden und von der zugehörigen Tabelle abhängen, werden Platzhalter verwendet. Man kann Eigenschaften für Mappings, Operatoren, Gruppen und Attribute definieren. Analog zum Oracle Warehouse Builder sind nur die Eigenschaften zu beschreiben, die nicht automatisch erzeugt werden können. Wenn man beispielsweise im Oracle Warehouse Builder einen Tabellen-Operator mit einem Expression-Operator verbindet, werden die Input-Attribute des Expression-Operators automatisch mit den richtigen Datentypen generiert. Der OWB Mapping Generator arbeitet auf die gleiche Weise. Die Property-Tabelle kann je nach Anforderung oder Komplexität der umzusetzenden Logik sowohl statische als auch dynamische Werte enthalten (siehe Listing 1 bis 3). Eine Eigenschaft kann mithilfe vordefinierter dynamischer Parameter festgelegt werden: Damit beschreibt man alle Attribute eines Operators zusammen, und nicht jedes Attribut einzeln. Erfordert die fachliche Logik eine umfassendere Berechnung der Werte, etwa abhängig vom Primärschlüssel der Tabelle oder von Datentypen der Attribute, bietet der OWB Mapping Generator die Möglichkeit, eine benutzerdefinierte Funktion anzulegen, die dann in der Property-Tabelle verwendet werden kann.

Da die Generierung der Mappings später mithilfe eines TCL-Skripts (OMB Plus) erfolgt, sind für die Definition von Templates minimale OMB-Plus-Kenntnisse erforderlich. Um komplexe Logiken mit den benutzerdefinierten Funktionen abbilden zu können, sind tiefere TCL-Kenntnisse notwendig.

Ohne Namenskonventionen läuft nichts

Da die Definition der Templates generisch erfolgt, braucht man nun einen Weg, diese mit den erforderlichen OWB-

Objekten (Module, Tabellen etc.) zu verbinden. Um die Generierung von einzelnen Mappings entsprechend seiner Anforderungen zu ermöglichen, kann man im OWB Mapping Generator Namenskonventionen und Umgebungs-Informationen ablegen. Dabei spielt der Begriff „Tabellenstamm“ (table radical) eine zentrale Rolle. Damit ist der gemeinsame Teil der Tabellennamen über alle im Mapping verwendeten Module hinweg gemeint (siehe Listing 4).

Der Tabellenstamm wird bei der Generierung von Mappings verwendet, um zusammengehörende Objekte in einem Mapping zu verbinden. Die Funktionsweise des Frameworks basiert auf der Annahme, dass alle verwendeten Datenbank-Objekte einer allgemeinen Namenskonvention folgen. In den bereitgestellten Konventionstabellen beschreibt man mithilfe der regulären Ausdrücke Namenskonventionen der Datenbank-Objekte innerhalb der OWB-Module und legt die Namenskonvention für die zu erzeugenden Mappings fest. Durch die einfache Erweiterbarkeit und Individualisierung des Frameworks können im OWB Mapping Generator beliebige Namenskonventionen abgebildet werden.

PROPERTY_NAME:	LOADING_TYPE
PROPERTY_VALUE:	INSERT/UPDATE

Listing 1: Operatoreigenschaft: statischer Wert

ATTRIBUTE_NAME:	$attr_name
PROPERTY_NAME:	EXPRESSION
PROPERTY_VALUE:	INGRP1.$attr_name

Listing 2: Attributeigenschaft: dynamischer Wert

PROPERTY_NAME:	SPLIT_CONDITION
PROPERTY_VALUE:	$func_get_scd2_close_set_cond

Listing 3: Gruppeneigenschaft: benutzerdefinierte Funktion

OWB Modul:	SOURCE	STAGE	CORE
Tabelle:	SRC_PRODUCT	STG_PRODUCT	PRODUCT

Listing 4: Tabellenstamm „PRODUCT“ in den Modulen „Source“, „Stage“ und „Core“

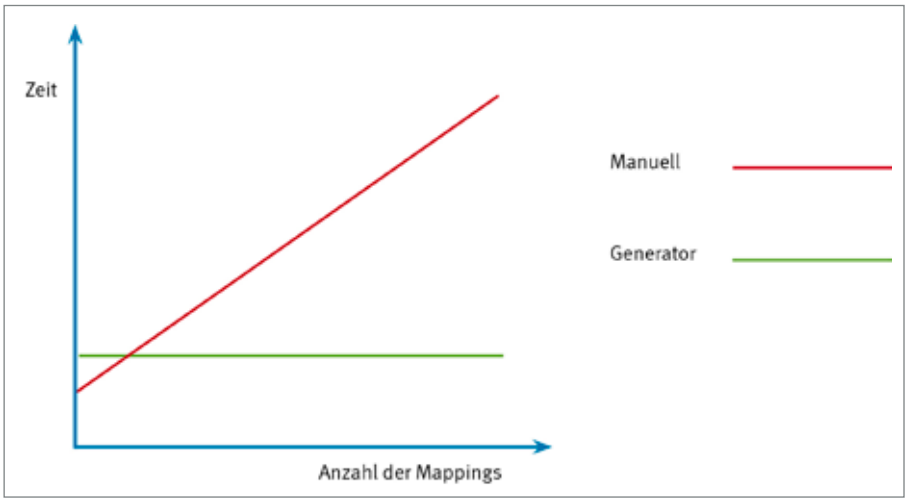


Abbildung 3: Zeitgewinn innerhalb aller Projektphasen beim Einsatz des OWB Mapping Generators

Fazit

Der Mapping Generator ist ein speziell entwickeltes Framework, mit dem man die Entwicklung in einem OWB-Projekt „industrialisieren“ und damit den Fertigstellungsprozess eines Data Warehouse enorm beschleunigen kann. Da die Definition der Templates exakt der Struktur von OWB Mappings folgt, ist eine Einarbeitung in das Framework sehr schnell möglich.

Das Framework hat nicht den Anspruch, den Oracle Warehouse Builder zu ersetzen, kann aber als Ergänzung bei vielen Aufgabenstellungen sehr hilfreich sein. Wie Abbildung 3 zeigt, lohnt sich der Einsatz des OWB Mapping Generators bei steigender Anzahl der Mappings. Man profitiert dabei in allen Projektphasen:

- Die Entwicklung wird beschleunigt, da statt einer großen Menge von Mappings nur noch ein Template designt werden muss.
- Vereinheitlichung und damit auch Qualitätsverbesserung des Codes ist ein unstrittiger Gewinn, den man in großen Projekten kaum noch erreichen kann.
- Der Testaufwand wird dank des generischen Ansatzes ebenfalls deutlich reduziert.
- Auf Neuanforderungen kann sehr schnell reagiert und ein Data Warehouse in kurzer Zeit neu aufgebaut werden.
- Die Wartungsaufwände werden bei der Verwendung des OWB Mapping Generators deutlich minimiert, wodurch man mehr Zeit für konzeptionelle Aufgaben und neue Projekte gewinnt.

Irina Gotlibovych
irina.gotlibovych@mt-ag.com

Business Intelligence nimmt einen hohen Stellenwert in der heutigen Unternehmenswelt ein. Dabei hat sich der Rahmen in den letzten Jahren deutlich erweitert. Neue Anwendungsgebiete wie Abrechnungslösungen im Rahmen der Operationalisierung im BI oder neue Einsatzbereiche wie „Real- & Near-Time"-BI-Lösungen in der Produktion sowie größere Datenmengen erhöhen die Komplexität der Anforderungen und sind zunehmend an besondere Vorgaben wie Revisionssicherheit oder Service Level Agreements geknüpft.

DWH/BI-Framework und Vorgehensmodell: der Weg zur BI Excellence

Alexander Neumann, arvato IT services GmbH

Zur Bewältigung dieser Herausforderungen reicht es nicht mehr aus, die Projekte fokussiert auf Time, Budget und Quality erfolgreich umzusetzen. Den gestiegenen Ansprüchen der Kunden können nur exzellente Lösungen gerecht werden.

Zur Erreichung der BI Excellence ist ein in der Organisation integriertes, ganzheitliches Realisierungsvorgehen bezüglich der zu erstellenden Produkte beziehungsweise produktnaher (BI-) Lösungen, Produkt-Realisierungsprozesse, Projektmanagement-Prozesse sowie des Projektteams erforderlich. Das Unternehmen des Autors steht täglich solchen Herausforderungen gegenüber.

Als interner IT-Dienstleister der arvato AG, ein Unternehmensbereich der Bertelsmann AG, berät arvato IT services die verschiedenen Markteinheiten und stellt durch integrierbare IT-Lösungen die Wettbewerbs- und Lieferfähigkeit der Markteinheiten im Sinne der Strategie „From Product to Solution" sicher. Dabei decken die Leistungen die gesamte Wertschöpfungskette der Markteinheiten ab und umfassen Beratung, Sourcing, Setup und Betrieb. arvato ist ein bedeutender Dienstleister für das Business Process Outsourcing (BPO). Integraler Bestandteil dafür ist die IT.

arvato IT services ist in technische (wie Business Intelligence oder Java-Entwicklung) und fachliche (wie Customer Relationship Management) Competence Center organisiert. Daneben übernehmen zentrale Services übergreifende Aufgaben (wie Strategie, Risk & Compliance) und dienen als Regulativ zwischen den Competence Centern (wie Portfoliomanagement).

Realisierung von BI-Projekten

BI-Projekte beziehen sich in diesem Kontext weitgehend auf die diversen BPO-Kundenlösungen und nicht auf die internen DWH-Lösungen. Diese haben ihre besonderen Herausforderungen oft in vielseitigen inhaltlichen und hohen Anforderungen in „Time to Market"-Aspekten. Neben den klassischen DWH/BI-Lösungen gehören revisionssichere B2B-Abrechnungslösungen sowie anspruchsvolle Lösungen im Fertigungsbereich zum regulären Projektgeschäft. Die angestrebte Excellence wird dabei durch eine Qualitätsmanagement-orientierte, ganzheitliche Betrachtungsweise der folgenden Komponenten sichergestellt:

- *Produkte*
  Ausgereiftes Portfolio aus produktnahen Lösungen, die vom klassischen DWH über komplexe Abrechnungslösungen bis zur „Real- und Near-Time"-Überwachung und Analyse von Geschäfts- und Fertigungsprozessen reichen und miteinander kombinierbar sind (siehe Abbildung 1)
- *Produktrealisierungs-Prozesse*
  Ein effizienter Mix aus spezialisierten Datenintegrationstools, Vorgehensmodell und technischem DWH/BI-Framework zur Realisierung von Portfolio-Komponenten
- *Projektmanagement*
  Innerhalb der Organisation verankerte Projekt- und Projektmanagement-Kultur (zentrale Steuerung über Project Management Office innerhalb des Competence Centers Quality Assurance Management (QAM), Projektabwicklung entsprechend den IPMA-Standards)
- *People (Team)*
  Qualifizierte und erfahrene Mitarbeiter, die verschiedene, im Vorgehensmodell definierte Rollen in Projekten einnehmen und über entsprechende Entwicklungspfade weitergebildet und zertifiziert werden

Im Folgenden werden die Produktrealisierungsprozesse (Schwerpunkt: Vorgehensmodell und DWH/BI-Framework) detailliert erläutert.

Produktrealisierungs-Prozesse von BI-Projekten

Bei den Realisierungs-Prozessen zur Implementierung von BI-Lösungen kommen neben den spezialisierten Datenintegrations-Tools ein DWH/BI-Framework sowie ein Vorgehensmodell zum Einsatz (siehe Abbildung 1).

Die spezialisierten Datenintegrations-Tools sind in der Regel nur bei der Umsetzung von Kernfunktionalitäten von DWH/BI-Anforderungen ausreichend. Wiederkehrende Anforderungen, die während des Realisierungsprozesses beziehungsweise im Betrieb von DWH/BI-Lösungen auftreten, werden in der Regel funktionell nur unzureichend unterstützt, beispielsweise automatisierte Prozesse für:

- Standardisierung der ETL-Entwicklung
- Deployment
- Ausführungsmonitoring
- Bereinigung
- Dokumentationsgenerierung

Wissen. Austausch. Erfolg.

Die Oracle-Konferenz

DOAG 2012 Konferenz + Ausstellung
20. – 22. November 2012
NürnbergConvention Center

- 400 Fachvorträge –
  Für jeden Anwender die besten Themen
- Top-Keynotespeaker: Sascha Lobo,
  Andrew Mendelsohn, Loïc le Guisquet, ...
- Networking und Erfahrungsaustausch:
  DBA's, Developer, Infrastruktur-Experten
  und Manager
- Wissen vertiefen:
  DOAG Schulungstag am 23. November

http://2012.doag.org

2012 DOAG Konferenz + Ausstellung

Eine Veranstaltung der DOAG mit

ORACLE iJUG Verbund SOUG AOUG MySQL

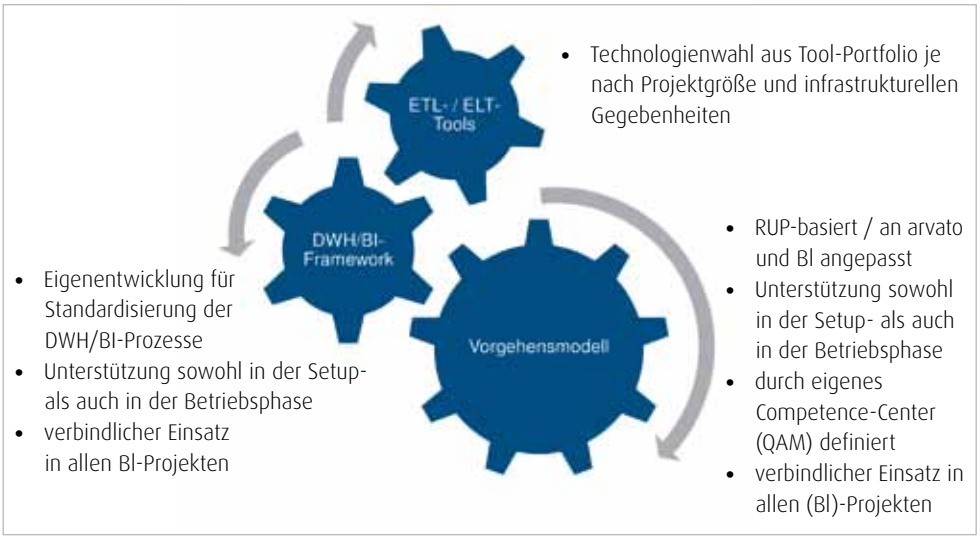


Abbildung 1: Effizienter Mix aus spezialisierten Datenintegrations-Tools, DWH/BI-Framework und Vorgehensmodell

Diese Defizite werden durch das DWH/BI-Framework, eine Eigenentwicklung für die Standardisierung des DWH/BI-Prozesses, eliminiert. Die Verwendung des DWH/BI-Frameworks wird wiederum durch ein speziell für DWH/BI-Belange adaptiertes Vorgehensmodell (RUP-basiert) – sowohl in der Setup- als auch in der Betriebsphase – festgelegt.

DWH/BI-Framework

Das Competence Center Business Intelligence hat ein DWH/BI-Framework entwickelt, das durch funktionelle Erweiterungen sowie Standardisierung von Entwicklungsprozessen (wie das Generieren kompletter Layer) die Umsetzung komplexer Kunden-Anforderungen zeitnah ermöglicht.

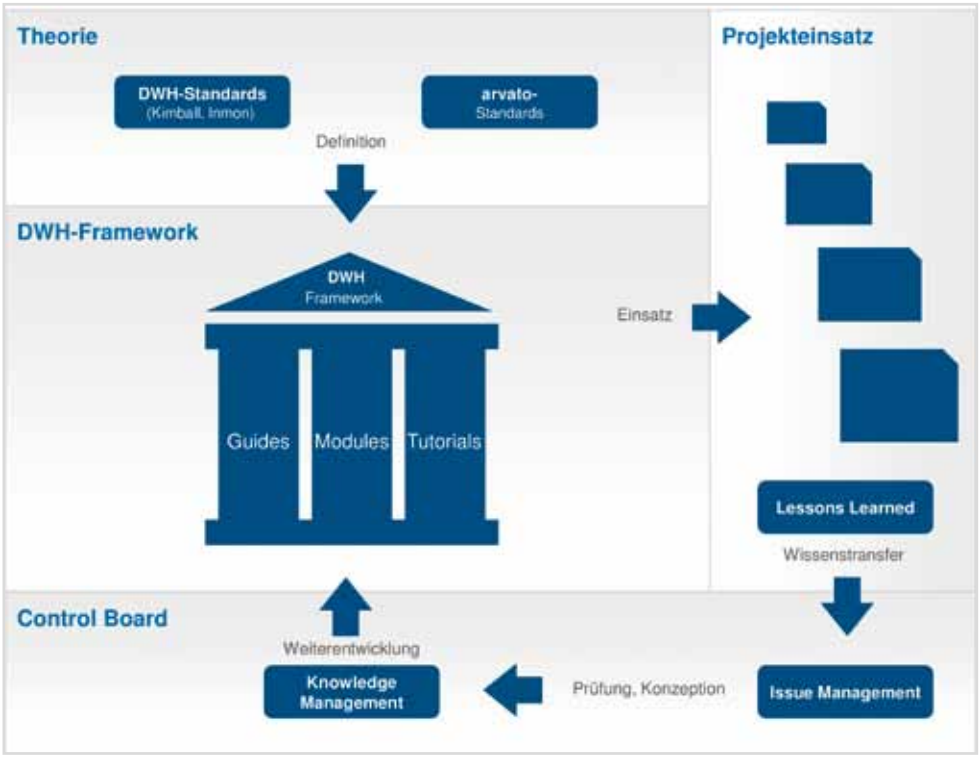


Abbildung 2: DWH/BI-Framework

Das Framework basiert auf etablierten BI-Paradigmen (Inmon und Kimball) sowie auf eigenen Standards und wird durch den verbindlichen Einsatz in allen BI-Projekten und den entsprechenden Erfahrungsrückfluss gelebt und ständig weiterentwickelt (siehe Abbildung 2). Neben der Vorgabe bezüglich der Leit-Architektur in BI-Projekten gibt das Framework konkrete Hilfestellungen bei der Umsetzung von Kunden-Anforderungen in Form von Guides (wie Vorgaben bezüglich des Designs), standardisierte Module zur Unterstützung sämtlicher Bereiche innerhalb der Projektrealisierung (Entwicklung, Testen, Dokumentation) sowie Tutorials als Anleitungen für konkrete Anwendungsfälle (siehe Tabelle 1). Die Framework-Verantwortung im Competence Center Business Intelligence ist durch eine Querschnittsfunktion organisatorisch geregelt.

Vorgehensmodell

Als Grundlage des zentralen, in der Organisation verankerten Vorgehensmodells für die Software-Entwicklung dient eine Adaptierung des Rational Unified Process (RUP) [2]. Weiterentwicklung und Überwachung des abteilungsübergreifenden Einsatzes des Vorgehensmodells übernimmt dabei das Competence Center Quality Assurance Management. Im Competence Center Business Intelligence wurde diese Adaption zusätzlich auf BI-spezifische Anforderungen zugeschnitten.

Das adaptierte Vorgehensmodell legt ein inkrementelles und iteratives Vorgehen fest, ist in Phasen (im Original: Disziplinen) gegliedert und Use-Case-basiert. Das Vorgehensmodell definiert Rollen (für die Ausführung der Tätigkeiten) sowie Artefakte (Dokumente zur Beschreibung der Anforderungen und realisierten Komponenten) im Rahmen der Umsetzung von BI-Projekten (siehe Tabelle 2).

Durch die einheitliche Begriffswelt innerhalb des Vorgehensmodells wird die Kommunikation sowohl innerhalb des eigenen Competence Centers als auch Competence-Center-übergreifend vereinfacht. Ein einheitliches Rollenverständnis ermöglicht standardisierte, rollenbasierte Mitarbeiterentwicklung.

Dokumentiertes, nachvollziehbares Vorgehen ermöglicht sowohl die Darstellung für Außenstehende als auch schnellere Einarbeitung neuer Mitarbeiter. Dabei wird die Kreativität nicht behindert – diese kommt bei der Lösungsfindung für den Kunden, bei Architektur und neuen Technologien zum Tragen – während immer wiederkehrende Routineprozesse hochgradig standardisiert ablaufen. Die sich daraus ergebenden Vorteile sind:

- Steigerung der Produktivität
- Verbesserung der Qualität
- Erleichterung der Führbarkeit von Projekten
- Reduktion von Risiken
- Verkürzung der Entwicklungszeiten durch schnelleres Projekt-Staffing

Bei der Realisierung von Projekten (etwa bei der Umsetzung eines Kundenbindungssystems) innerhalb von arvato IT services kommt das Vorgehensmodell in allen am Projekt beteiligten Einheiten zum Einsatz. Die fachlichen Competence Center (wie Customer Relationship Management und Customer Service Management) beraten die Markteinheiten hinsichtlich geeigneter Kunden-Lösungen. Die IT-seitige Umsetzung steuern die technischen Competence Center bei, darunter Loyalty Management, BI, Customer Intelligence Services, Service Management. Das übergreifende Qualitätsmanagement wird durch das Competence Center QAM sichergestellt. Dabei werden die beteiligten Markteinheiten (wie Print & Lettershop, Logistics, Financial Services) ebenso eingebunden wie auch externe Komponenten wie Endkundensysteme (Debitorenmanagement, Warenwirtschaftssysteme, Kassensysteme).

Guides (Auszug): Design-Guide Configuration-Guide Performance-Guide Partitioning-Guide	**Module (Auszug):** Logging Automatisierte Partitionierung ETL-Dokumentation Qualitätssicherung Prozessfluss-Steuerung Release-Notes-Generierung Mapping- (Paket-) Generierung Prozessfluss-Generierung
Tutorials (Auszug): Vorgehen bei Prozessabbrüchen Software-Installation Generierung von Mappings/Paketen Einrichtung Repository Migrationspfade	

Tabelle 1: Bestandteile des DWH/BI-Frameworks

Phasen: Acquisition Business Modeling Requirements Analysis and Design Implementation Test Deployment Configuration and Change Mgmt. Project Management Environment Operations and Support	**Meta Use Cases (Auszug):** DWH-Projekt initiieren Kostenschätzung erstellen Systemanforderungsspezifikation erstellen Infrastruktur einrichten Anwendung entwerfen System implementieren Projektfortschritt überwachen Testfallspezifikation erstellen Integrationstest begleiten DWH Release abschließen Deployment durchführen Change Request erstellen Hotfix durchführen
Rollen (Auszug): Administrator (DB) Administrator (DWH) Data-Warehouse-Architekt Entwickler Projektleiter IT-Gesamtprojektleiter Ressourcenmanager System Analyst Servicemanager Test Manager Tester Testdesigner	**Artefakte (Auszug):** Systemanforderungsspezifikation Schnittstellen-Spezifikationen Reportspezifikation Würfelspezifikation ETL-Konzept BI-Konzept Release-Notes Change Request Wartungshandbuch Wartungslogbuch Testplan Testkonzept Testabschlussbericht

Tabelle 2: Bestandteile des adaptierten Vorgehensmodells

Zusammenspiel aller Komponenten

Das (DWH/BI-)Framework ist im Vorgehensmodell integriert. Dabei kommen je nach Phase verschiedene Komponenten des Frameworks zum Einsatz und generieren den erforderlichen Output in deutlich schnelleren Realisierungszeiten (als in der Vergangenheit in Projekten ohne Framework-Einsatz). Im Übrigen verfügen alle technischen Competence Center über entsprechende Frameworks. Abbildung 3 zeigt das Zusammenspiel des DWH/BI-Frameworks und des Vorgehensmodells visualisiert. Die Einsparungen in Prozent gegenüber den Projekten ohne Framework-Einsatz sind grün gekennzeichnet. Abbildung 4 stellt am Beispiel der Phase „Implementation" die Integration des DWH/BI-Frameworks innerhalb des Vorgehensmodells dar.

Als Ergebnis der Phase „Analysis & Design" bildet das ETL-Konzept den Ausgangspunkt für die Phase „Implementation". Im ETL-Konzept werden sämtliche Vorgaben bezüglich der Backend-seitigen Implementierung festgelegt. Die Umsetzung der ETL-Prozesse erfolgt – je nach Komplexität – unter Verwendung verschiedener Framework-Module, etwa bei der Generierung kompletter vorgelagerter Layer (beispielsweise eines Operational Data Stores). Die projektseitige Konfiguration und Steuerung der Module erfolgt über ein Eclipse-Frontend. Die entsprechenden Modul-Sourcen und -Bi-

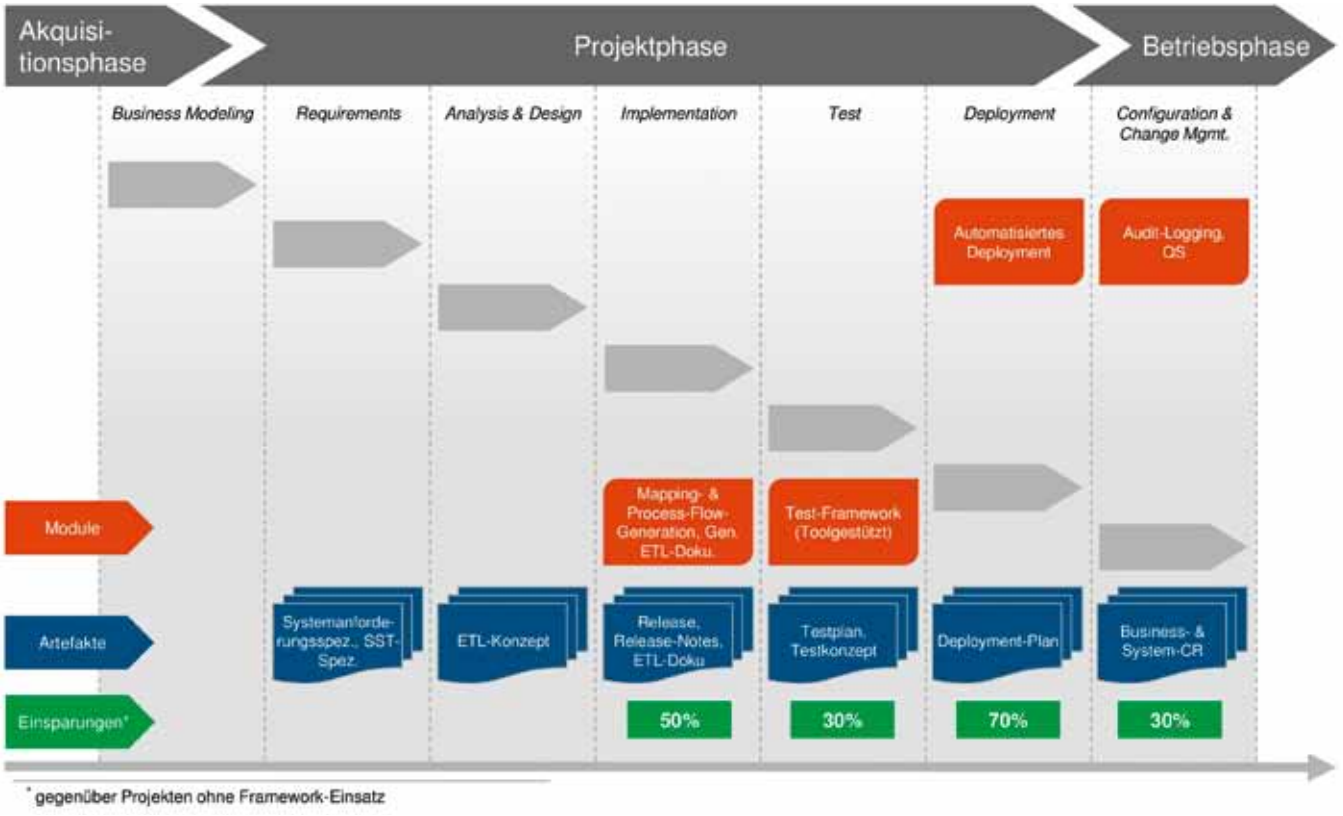


Abbildung 3: Zusammenspiel von Vorgehensmodell und DWH/BI-Framework

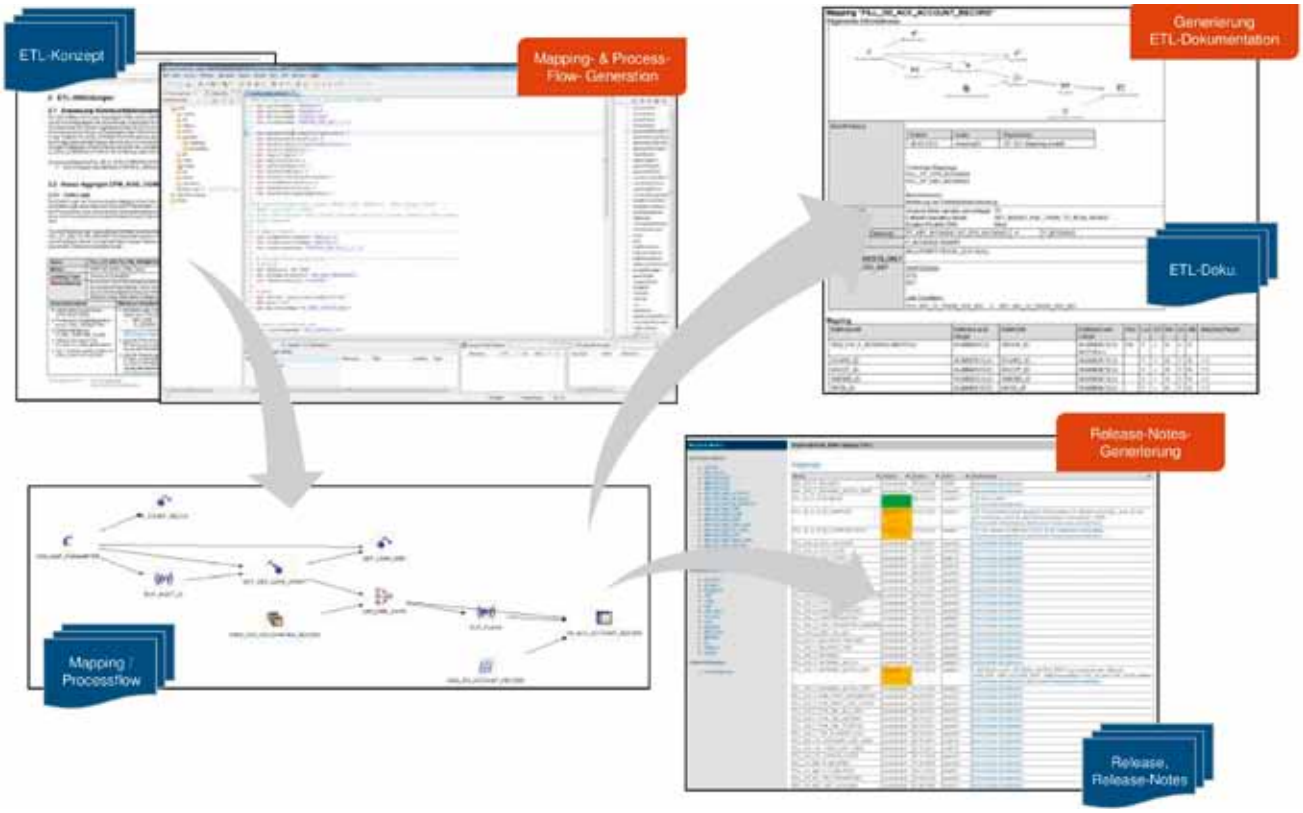


Abbildung 4: Zusammenspiel von Vorgehensmodell und DWH/BI-Framework am Beispiel der Phase „Implementation“ [1]

bliotheken sind zentral in einem Versionierungstool (SVN) verwaltet. Am Beispiel eines Projekts unter Verwendung von Oracle-Technologien (OWB als ELT-Tool) wird der Layer Operational Data Store zu 100 Prozent automatisiert über „OMB+“-Komponenten generiert. Dabei werden Mappings verschiedener Typen und Prozessflüsse automatisiert angelegt und in verschiedenen Umgebungen ebenfalls automatisiert eingesetzt. Darüber hinaus erzeugen Framework-Module automatisiert eine technische Dokumentation (Word-Format) sowie Release-Notes auf

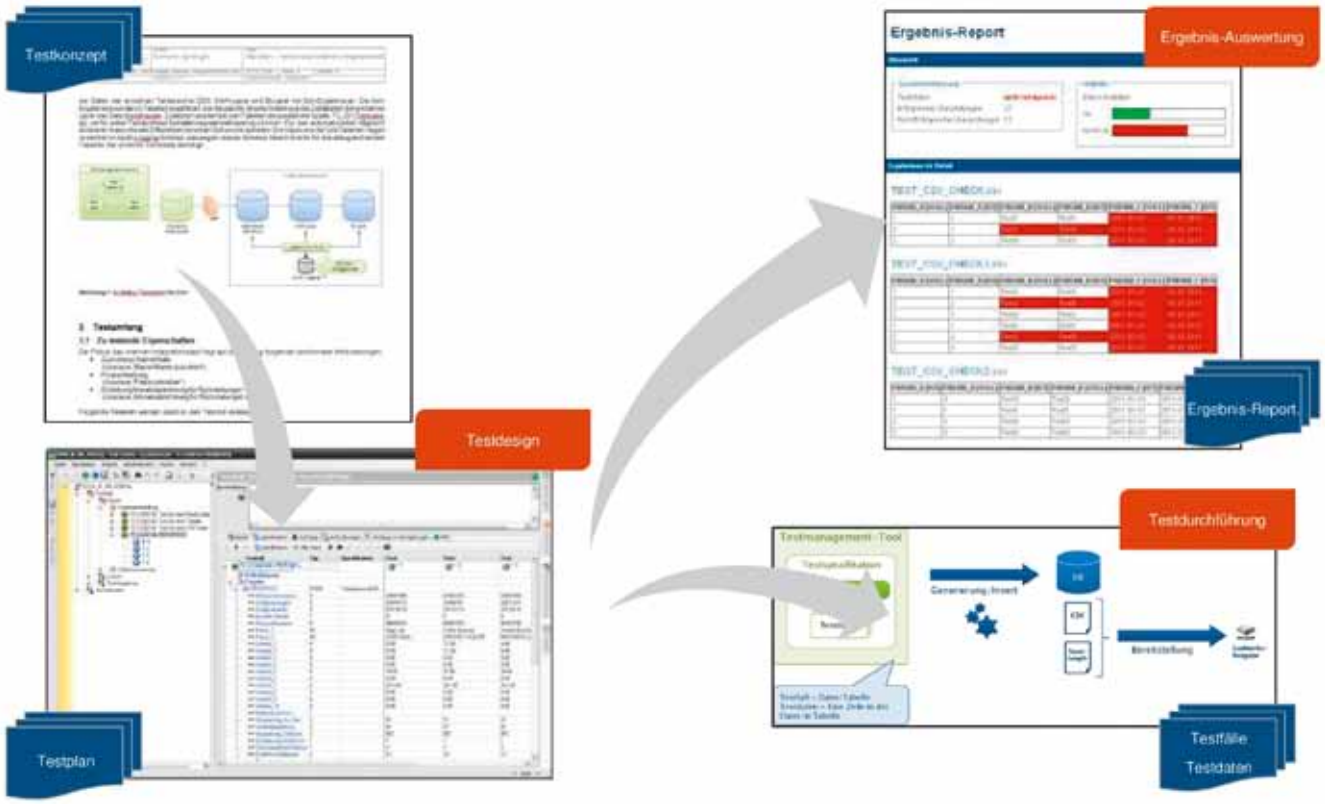


Abbildung 5: Zusammenspiel von Vorgehensmodell und DWH/BI-Framework am Beispiel der Phase „Test“

Attribut-Ebene (HTML-Format). In analoger Weise gestaltet sich die Integration des DWH/BI-Frameworks innerhalb des Vorgehensmodells im Rahmen der Phase „Test“ (siehe Abbildung 5).

Das Vorgehen innerhalb der Phase „Test“ ist folgendermaßen definiert: Im Rahmen des Test-Designs wird je nach Anwendungsfall vom DWH/BI-Modultest über den systemübergreifenden Integrationstest mit allen beteiligten Schnittstellen bis hin zum Kunden-Abnahmetest die entsprechende Testkonzeption definiert. Diese umfasst die zeitliche Planung der Testdurchführung, erforderliche Ressourcen sowie entsprechende Testfälle im Rahmen der White-Box-Testmethode. Die zugehörigen, durch das Vorgehensmodell vorgegebenen Artefakte innerhalb der Testkonzeption sind:

- Testkonzept
- Testplan
- Testszenarien
- Testfälle
- Testdaten

Im Rahmen der anschließenden Testdurchführung und Testauswertung kommen – analog zur Phase „Implementierung“ – weitere Bestandteile des DWH/BI-Frameworks zum Einsatz: Simulation der Quellsysteme, automatisierte Bereitstellung der Testdaten, ebenfalls automatisierte Ausführung der betroffenen (DWH/BI-)Komponenten sowie die anschließende Ergebnis-Protokollierung im Rahmen der Testdurchführung. Innerhalb der Testauswertung werden die Ergebnisse Framework-gestützt überprüft (etwa durch einen automatisierten Soll-Ist-Vergleich) und bei Bedarf zur Nachverfolgung übergeben [2]. Entsprechende Integration des DWH/BI-Frameworks innerhalb des Vorgehensmodells findet in den Phasen „Deployment“ sowie „Configuration & Change Management“ statt.

Fazit

Die Produkt-Realisierungsprozesse sind geprägt durch hohe Standards sowie deren Verankerung in der Organisation. Der verbindliche Einsatz beider Komponenten wird überwacht, alle Projekte müssen sich in regelmäßigen Abständen Vorgehensmodell- und Framework-Reviews unterziehen). Das Vorgehensmodell stellt abteilungsübergreifende Homogenität der Realisierungsprozesse sicher. Das DWH/BI-Framework ist hochgradig standardisiert und dabei flexibel, erweiterbar, wartbar sowie einfach zu handhaben. Dies ermöglicht eine erhebliche Reduktion der Projektaufwände (im Setup und Betrieb) sowie qualitativ hochwertige Lösungen trotz schnellerer Realisierungszeiten.

Referenzen

[1] Alexander Neumann, Dominik Sprenger (2012): DWH/BI-Framework und Vorgehensmodell, Vortrag auf der DOAG 2012 Business Intelligence Konferenz

[2] Philippe Kruchten (2003): Rational Unified Process 3rd Edition: An Introduction

[3] Dominik Sprenger (2009): Konzept für automatisierte Tests im Data-Warehouse-Umfeld.

Alexander Neumann
alexander.neumann@bertelsmann.de

Das Business Intelligence Competency Center (BICC) ist eine sehr gut geeignete Organisationsform, um BI-Ressourcen eines IT-Dienstleisters zusammenzubringen, die über Geschäftsbereiche und regionale Geschäftsstellen verteilt sind. Dadurch lassen sich die Qualität und der Umfang der Dienstleistungen im Bereich Data Warehousing (DWH), Business Intelligence (BI) und darüber hinaus enorm steigern; gleichzeitig werden Wissenstransfer und Technologie-Know-how gefördert.

To BICC or not to be – auch für einen IT-Dienstleister

Manfred Dubrow, Robotron Datenbank-Software GmbH

Der Artikel beschäftigt sich mit der Motivation zur Etablierung eines BICC und mit den zu erreichenden Zielen am konkreten Beispiel des Robotron-Verbunds, eines spezialisierten Unternehmens für Lösungen auf Basis der Oracle-Technologie. Er zeigt, wie typische Eigenschaften und Aufgaben eines BICC auf die gewünschte Organisationsform angewendet werden können.

Mehrheitlich wird ein BICC diskutiert, wenn in Organisationen, die BI-Werkzeuge und -Verfahren einsetzen, IT (als BI-Betreiber) und BI-Fachanwender effektiv kooperieren sollen. Robotron ist jedoch hauptsächlich BI-Dienstleister und nur sekundär auch selbst BI-Anwender. Es stellt sich die Frage, ob es da sinnvoll ist, die Prinzipien und Praktiken eines BICC an die Belange eines IT-Dienstleisters zu adaptieren und damit Qualität, Effektivität und Effizienz der BI/DWH-Vorhaben für Kunden zu entwickeln und nachhaltig zu sichern.

Einige Faktoren, die bei einem BI/DWH-Betreiber ein BICC gewöhnlich motivieren, sind für einen BI/DWH-Dienstleister kaum relevant (wie Interessenbündelung von Fachbereichen und IT oder Etablierung von Daten-, Prozess- und Kennzahlenstandards). Andererseits gibt es gute Gründe für den Unterhalt einer spezifischen Organisationsform für BI-Vorhaben.

Business-Intelligence-Lösungen sind für Robotron ein strategisches Geschäftsfeld. Das Wachstumspotenzial bei BI/DWH, speziell der immer breitere Einsatz bei analytischen und entscheidungsstützenden Tätigkeiten, führt dazu, dass einschlägiges Personal rekrutiert, neue Lösungs- und Anwendungsbereiche erschlossen, das BI/DWH-Leistungsangebot gestärkt, externe Kooperationen angestrebt und BI/DWH zunehmend in komplexere Anwendungslösungen integriert werden.

Die BI/DWH-Fähigkeiten von Robotrons Gesellschaften und Geschäftsstellen in Deutschland, der Schweiz und in Österreich sind unterschiedlich ausgeprägt, obwohl die BI-Anforderungen in den jeweiligen Regionen stetig wachsen. Die Möglichkeiten müssen demnach optimal und koordiniert entwickelt werden, damit BI/DWH-Leistungen auf allen bearbeiteten Märkten angeboten werden (können), BI/DWH-Fähigkeiten optimal in Projekten Eingang finden und geeignete Ressourcen untereinander austauschbar sind. Das gilt ebenso bei voneinander isolierten Geschäftsbereichen, damit BI/DWH-Wissen und -Nutzen im Hinblick auf bessere Produkte und Leistungen weitgehend einheitlich und breiter angewendet werden.

Nicht zuletzt wird ganzheitliche BI/DWH-Beratung nachgefragt. BI/DWH-Projekte haben meist zusätzlich eine planerische und organisatorisch/strategische Komponente, die entsprechenden Beratungsbedarf hinsichtlich Strategie, Methodik, Technik und Betrieb erzeugt. Dem sollte sich ein BI/DWH-Leistungsanbieter gewachsen zeigen können.

Von der Vision zu Maßnahmen

Zunächst gilt es, eine Vision davon zu entwickeln, welchen Stellenwert Business Intelligence und Data Warehousing im Dienstleistungsangebot haben, genauer: ob BI/DWH ein strategisches Geschäftsfeld sein soll. Strategisch heißt, ein Ziel unter Berücksichtigung der verfügbaren Mittel und Ressourcen längerfristig und planvoll anzustreben. Bevor man sich auf diesen Weg begibt, sollte mit einer schonungslosen Analyse von Stärken und Schwächen begonnen werden. Im Ergebnis lassen sich sehr gut Maßnahmen ableiten, insbesondere wie die herausgearbeiteten Stärken und Chancen genutzt werden können, um BI/DWH als strategisches Geschäftsfeld zu etablieren.

Für das gleichzeitige Betrachten von Stärken, Schwächen, Möglichkeiten und Gefahren (Risiken) ist die SWOT-Analyse weit verbreitet. Sie geht auf Albert S. Humphrey zurück und steht für Stärke (Strength), Schwäche (Weakness), Chance (Opportunity) und Risiko (Threat). Damit lassen sich sowohl eine Sicht nach innen (Stärken und Schwächen) als auch auf das Marktumfeld (Chancen und Risiken) herstellen und diese miteinander kombinieren. Das Hauptaugenmerk liegt auf der S-O-Strategie, also eigene Stärken einzusetzen, um Chancen zu ergreifen. Daneben sollte aber auch betrachtet werden, wie Chancen trotz bestehender Schwächen nicht verpasst werden, wie sich Risiken durch Stärke bewältigen lassen oder wie Schwächen zu mindern und Risiken zu meiden sind. Im nächsten Schritt werden Ziele definiert und durch Maßnahmen unterstützt. Im konkreten Fall sind dies:

- Bündelung und Koordinierung der BI/DWH-affinen Ressourcen über alle Bereiche und sonstigen Unternehmenseinheiten; außerdem Ge-

winnung von mehr personeller Flexibilität, Nutzung der bestehenden personellen Plattform für deren kontinuierlichen Ausbau und Etablierung von einheitlich verwendeten Standards (Methoden und Praktiken).
- Alleinstellendes BI/DWH-Lösungs- und Leistungsangebot sowie entsprechend hohe Beratungsqualität an allen Standorten unter Nutzung des jeweils vorhandenen Kunden- und Vertriebspotenzials.
- BI ist (fast) überall. Umfassende Nutzung von BI-Methoden und -Techniken für jegliche Datenanalyse und Informationsgewinnung. Nutzung des BI/DWH-Potenzials für die Integration in Produkte und Lösungen.
- Schaffung eines ganzheitlichen Angebots im BI/DWH-Umfeld von fachlicher und methodischer Beratung bis hin zu Implementierung. Das Angebot soll sich über den jeweils gesamten aktuellen Oracle-BI/DWH-Produkt-Stack erstrecken, aber auch Leistungen mit Produkten anderer Hersteller umfassen und insgesamt mit der technologischen Entwicklung schritthalten.
- Generell bessere Vernetzung der Unternehmensstandorte durch Wissenstransfer sowie gemeinsame Markt- und Realisierungsaktivitäten in dem bestimmten Themensektor BI/DWH.

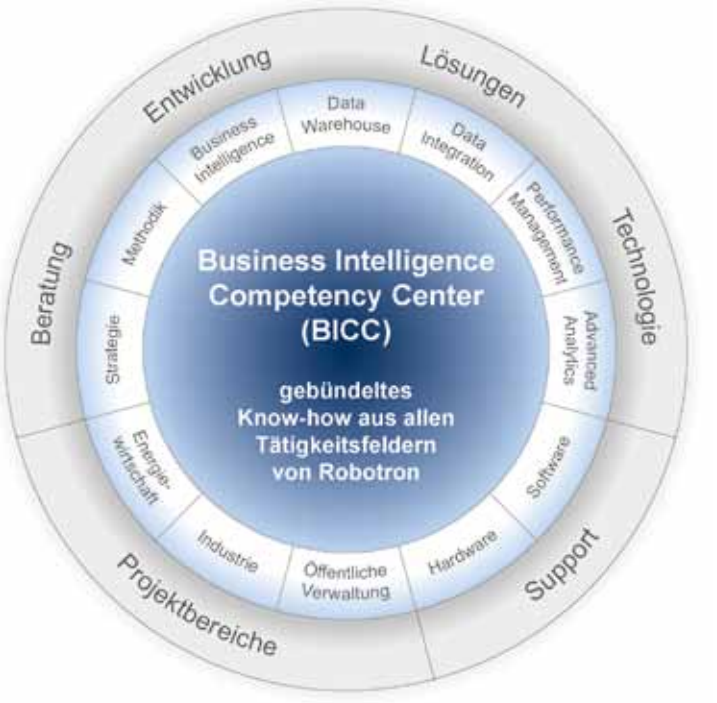


Abbildung 1: Aufgaben und Beteiligte bei der Umsetzung des BICC-Konzepts

Eignung eines BI Competency Centers

BICC ist ein gut erforschtes, anerkanntes und bewährtes Konzept zur Umsetzung einer BI-Strategie als Teil der Unternehmensstrategie. Für die programmatische und organisationale Operationalisierung der Strategie stehen Regelwerke bereits zur Verfügung. Ein BICC ist eine spezifische Form zum Aufbau und Wirken einer BI-Organisation im Rahmen der Umsetzung einer BI-Strategie. Diese Regeln gilt es zu adaptieren und zu implementieren. Ein BICC hat folgende Eigenschaften:

- Zentralisiert das BI-Wissen im Unternehmen
- Ist eine mögliche Organisationsform für eine BI-Governance (im Gegensatz zum Betrieb aus rein operativ-taktischer Notwendigkeit)
- Agiert organisationsweit, also fachbereichs- und projektübergreifend
- Fördert organisationsweite Rahmensetzung und Kooperation
- Kann mit geringer bis hoher fachlicher und technischer Tiefe angelegt sein
- Umfasst eine breite Wissenspalette (Business (Fachlichkeit), IT, Analytik) aus mehreren Organisationsbereichen

Das BICC ist demnach gut geeignet, die gestellten Ziele zu verfolgen und zu erreichen. Die Implementierung des BICC und seine strategische Positionierung sind eine zusätzliche Referenz für die BI/DWH-Fähigkeiten eines Unternehmens hinsichtlich Beratung und Implementierung von BI/DWH.

Aufgaben des BICC

Das Robotron BICC hat sehr vielfältige Aufgaben im Bereich BI/DWH (siehe unten). Die mit den Aufgaben verbundenen Leistungen und Ergebnisse richten sich primär nach innen, sie entstehen also nicht im direkten Kundenauftrag, sondern dienen der Erfüllung der Unternehmensziele. Gleichwohl sollen sie einen entscheidenden Wettbewerbsvorteil bei externen Dienstleistungen hervorbringen (siehe Abbildung 1).

Bündelung der BI-Ressourcen: Diese Aufgabe zielt auf die Errichtung einer Arbeitsgemeinschaft von Personen des gesamten Unternehmens, überwiegend zum Erfahrungsaustausch und jenseits des täglichen Projekt- und Entwicklungsgeschäfts:

- *Zusammenführung der BI/DWH-Ressourcen der Geschäftsbereiche, Geschäftsstellen und Landesgesellschaften als verteilt virtuelle Einheit*
  - Bereich Energie
    Unterstützung spezifischer Komponenten der hauseigenen Produkte für das Energiedatenmanagement (EDM, etwa eDWH, Analytik, Reporting)
  - Bereich Industrie
    Controlling der Energieeffizienz, Manufacturing Intelligence, BI/DWH in der Industrie
  - Bereich Öffentliche Verwaltung
    Statistik, Monitoring, Controlling mittels BI/DW
  - Bereich Support
    Technische Unterstützung bei Oracle Business Intelligence und Oracle-Data-Integration-Plattformen, Synergie von IT, Oracle-Pro-

dukten und BI/DWH-Fachanwendungen
  - Schulung (Oracle Approved Education Center)
    Gewinnung von Praxis-Know-how für BI/DWH-Schulungen
  - Robotron Austria/Schweiz
    Entwicklung von BI/DWH-Ressourcen für Projekte in Österreich beziehungsweise in der Schweiz
  - Geschäftsstellen in Deutschland
    Kundennahe DWH/BI-Dienstleistungen
- Koordinierung der Ressourcen für den effizienten Projekteinsatz
- Einbeziehung von Studierenden

Vorlaufprojekte für Technologieerprobung: Die Erprobung und Aneignung neuer Technologien ist eine wesentliche Aufgabe der Arbeitsgemeinschaft „Technologischer Vorlauf". Dafür werden quartalsweise interne Projekte geplant und durchgeführt:

- Interne BI/DWH-Projekte
  - Oracle BI Technologie Stack
  - Herausforderung Big Data
  - Prototypische Umsetzung bestimmter Anwendungsfälle
  - Eindringen in Oracle Fusion Middleware (OFM) Applikationen (analytische Komponenten)
- Entwicklung/Anwendung von Tools für ein effizientes und qualitätsgerechtes BI/DWH Project Lifecycle Management (PLM) als Vorstufe eines BI Application Lifecycle Management (ALM/BI)

Förderung von Innovation und BI-Anwendungslösungen: Neben den internen Technologieprojekten, in denen es primär um den Umgang mit den Oracle-Produkten geht, soll der Blick geweitet werden, indem Projekterfahrungen und der Diskursbereich BI/DWH insgesamt betrachtet werden. Im Ergebnis sollen Lösungswissen und neuartige Lösungen entstehen:

- Betreuung von studentischen Abschlussarbeiten zu Robotron-Themen, Begleitung von Promotionen
- Einsatz von Werkstudierenden, Praktikanten etc. in Untersuchungsaufgaben
- Kooperation mit BI/DWH-Lehrstühlen in Deutschland bei Zukunftsthemen durch Betreuung von wissenschaftlichen Abschlussarbeiten
- Konzeption und Entwicklung neuartiger Datenmanagement- und Analyseanwendungen mittels spezifischer Projekte
- Entwicklung von Robotron BI-Applikationen im Bereich Energie, Manufacturing und ÖV
- Umsetzung von bestimmten, intern entwickelten Anwendungskomponenten und Anwendungen für Kunden in innovativen BI/DWH-Lösungen, die von Dritten in einem bestimmten Anwendungskontext nutzbar sind

Sonstige Aufgaben sind:

- *BI/DWH-Beratung*
  - Unterstützung bei der Umsetzung von Anforderungen (Rat und Tat)
  - Ausbau der Beratungskompetenz im Bereich BI/DWH-Organisation (Strategie, Governance, BICC-Implementierung, Anwendungsintegration)
  - Anwendung des Rahmenwerkes „IT Strategies for Oracle".
- *Qualifizierung, Ausbildung*
  - Planung und Koordinierung der gezielten BI/DWH-Schulung des BICC-Personals (intern, extern, Web)
  - Unterstützung von Oracle-Spezialisierungen für BI und DW
- *Bevorratung von Best Practices*
  - Erarbeitung eines Arsenals an praktischen Fähigkeiten, Fertigkeiten und Lösungsmustern zur effektiven und effizienten Verwendung in BI/DWH-Vorhaben (Beratung, Konzeption, Implementierung, Betrieb, Schulung)
  - Entwicklung von BI-Standards und -Richtlinien
- *Projektleitung und Erarbeitung einer Projektdurchführungsmethodik*
  - Koordinierung von BI-Initiativen und BI-Projektmanagement
  - Entwicklung eines Vorgehensmodells für BI/DWH-Projekte mit Phasenstruktur, Zuordnung von V-Modell-XT-Entscheidungspunkten, Berücksichtigung der inkrementellen Entwicklung und der abschließenden Ergänzung von bestehenden Prozessbeschreibungen und Dokumentenvorlagen des unternehmensweiten QM-Systems
  - Evaluierung/Anwendung alternativer Methoden wie Agiles BI/DWH.
  - Fachliche Untersetzung eines toolgestützten PLM
- *Gremienmitwirkung, Öffentlichkeitsarbeit*
  - Aktive Beteiligung in einschlägigen BI/DWH-Gremien (DOAG, Oracle Partner Community, TDWI, BITKOM, Lehrstühle etc.) durch Gremientätigkeit und Vortragsangebote
- *Wissenstransfer*
  - Multiplikation von erlangtem Wissen, etwa aus Projektarbeit, Gremienmitwirkung, externer Schulung, Veranstaltungsteilnahme
  - Beobachtung geeigneter Informationskanäle und gezielte Informationsverteilung; Wissensmanagement
- *Unterstützung von Marketing und Vertrieb*
  - Erarbeitung und Aktualisierung von Unterlagen für die Kundenansprache und Publikationen
  - Mitwirkung bei der Anbahnung von BI/DWH-Vorhaben

Organisatorische Einbettung des BICC

Bereits bei der Etablierung eines BICC sind wichtige organisatorische Rahmenbedingungen zu klären. Robotron orientiert sich dabei an dem Oracle-Drei-Säulen-Modell (siehe Abbildung 2, Quelle: „The Business Intelligence Competency Center: Enabling Conti-

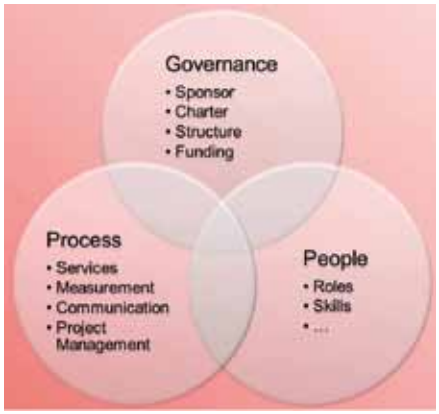


Abbildung 2: Oracle-BICC-Framework

nuous Improvement in Performance Management", An Oracle White Paper, January 2012). Die erste Säule befasst sich mit Führung und Struktur:

- Executive Sponsor ist ein Bereichsleiter, unterstützt von dem Leiter der Stabsstelle Technologie und Qualitätsmanagement (QM) und dem Leiter der auf BI/DWH spezialisierten Abteilung
- Das BICC agiert auf der Basis eines Grundsatzprogramms (Charter), in dem Ziele, Aufgaben und Organisation festgelegt sind
- Das BICC ist eine verteilt-virtuelle Einheit und als Unterarbeitsgruppe (UAG) in der Arbeitsgruppe Technologischer Vorlauf (AG TV) verankert. Die Mitglieder der UAG stammen aus dem BI/DWH-Kernteam und auf freiwilliger Basis aus weiteren Unternehmenseinheiten
- Dem BICC wird quartalsweise (Finanzquartal) auf Antrag über die AG TV ein Budget zugeteilt. Die Hauptverwendung ist die Durchführung interner Projekte

Die zweite Säule befasst sich mit dem Personal. Vordefinierte Rollen sowie die erwarteten beziehungsweise zu entwickelnden Fähigkeiten werden mit konkreten Personen besetzt. Aus der Spezifik des BICC ergibt sich, dass bestimmte Rollen unbesetzt bleiben können, etwa Personal, Technik, Betrieb.

Das Kernteam des BICC wird von der Projektabteilung gebildet und führt vorwiegend BI/DWH-Projekte in Industrie und öffentlicher Verwaltung durch. Korrespondenten des Kernteams sind bestimmte BICC-Mitglieder anderer Abteilungen und Organisationseinheiten, die eine oder mehrere der gewünschten Fähigkeiten aufweisen und eine der benötigten BI/DWH-Rollen ausfüllen. Zusätzlich sind bestimmte Mitarbeiter für Vertrieb/Marketing assoziiert. Ein leitender Systemberater des BI/DWH-Kernteams übernimmt die operative Leitung des BICC. Er handelt dabei wie ein Projektleiter und nutzt für Planung, Budgetierung und Controlling die dafür vorgesehenen Werkzeuge. Pro Geschäftsquartal wird ein Abschlussbericht gefertigt.

Backup & Recovery Oracle Datenbanken

Wir vermitteln Ihnen praktisches anwendbares Wissen!

Lernen Sie anhand eines Demo-Systems den sicheren Umgang mit den geeigneten Werkzeugen und den zahlreichen Möglichkeiten für die Sicherung und Wiederherstellung von Oracle Datenbanken. Das Ergebnis wird Ihnen ihren Arbeitsalltag erleichtern. Unser Workshop im Rahmen des Schulungstages der DOAG Konferenz 2012 erfolgt erfahrungsorientiert und praxisnah.

Ihr Referent:
Volker Volker Mach ist Leiter des Fachbereiches IT System Services der MT AG. Als Oracle Certified Professional kümmert er sich mit seinen 27 Mitarbeitern um Remote- und Vorort-Infrastrukturthemen der Hersteller Oracle, Microsoft, IBM, SAP Basis sowie OpenSource Technologien.

Sichern Sie sich jetzt Ihren Platz für den Schulungstag am 23.11.2012. Die Teilnahmegebühren für den Workshop betragen 490,- Euro zzgl. Mehrwertsteuer. Oder zu attraktiven Konditionen über den Konferenzbundle-Preis.

Weitere Infos sowie Tickets unter:
http://bit.ly/QJFbps

MT AG

Die dritte Säule für Prozesse und Abläufe, etwa für zu erbringende Dienstleistungen (hier gleichbedeutend mit den oben angeführten Aufgaben) und deren Bewertung, trifft für einen BI/DWH-Dienstleister eher nicht zu. Wichtig ist für ein BICC, das als virtuelle Einheit agiert, eine gute Kommunikationsplattform zu betreiben. Die Robotron-BICC-Teamsite auf der Basis des MS Sharepoint DMS ist hier das zentrale Kommunikationselement. Sie ist wie alle weiteren Fachteamsites in die Enterprise-Suche integriert und steht allen Mitarbeitern als Informationsquelle zur Verfügung. Der schreibende Zugriff bleibt (zunächst) den Teammitgliedern vorbehalten.

Fazit

Das vielfältige Aufgabenspektrum im Zusammenhang mit BI/DWH-Dienstleistungen für Kunden erfordert eine breite Palette an entsprechenden Fähigkeiten. Für ein einheitliches Handeln auf hohem fachlichen und technologischen Niveau ist die Koordinierung von Wissen und Fähigkeiten unabdingbar. Der Beitrag zeigte beispielhaft, wie sich ein BI/DWH-Dienstleister bei den anstehenden Aufgaben der Prinzipien und Verfahrensweisen eines BI Competency Centers bedienen kann. Erste Erfahrungen zeigen, dass ein BICC dafür durchaus sehr gut geeignet ist.

Manfred Dubrow
manfred.dubrow@robotron.de

In Zeiten weltweiter Vernetzung und gestiegener Anforderungen der Kunden müssen Unternehmen über alle Branchen hinweg zeitnah auf deren Bedürfnisse reagieren. Vor allem für die Fachabteilungen bedeutet dies, anhand von Zahlen, Daten und Fakten kurzfristig Entscheidungen zu treffen.

Mehr Unabhängigkeit, Flexibilität und Ergebnisorientierung mit Self-Service BI

Matthias Spieß, SHS VIVEON AG

Müssen erst Release-Zyklen in BI-Entwicklungen abgewartet werden, gehen wertvolle Chancen und Potenziale verloren. Self-Service Business Intelligence (SSBI) bietet den Anwendern hingegen die Möglichkeit, IT-unabhängige Analysen und Reports selbst zu generieren und gegebenenfalls mit weiteren Daten anzureichern. Dieser BI-Ansatz ist neben einer agilen Vorgehensweise in der Entwicklung eine weitere Komponente für die erfolgreiche Zukunft von Business Intelligence.

Die Anforderungen an ein professionelles Kundenmanagement sind in den letzten Jahren stetig gewachsen. Besonders die schnellen und permanenten Veränderungen der Geschäftsstrukturen und des Kundenverhaltens sowie eine hohe Dynamik der wirtschaftlichen Bedingungen stellen Unternehmen heute vor große Herausforderungen. In der Praxis zeigt sich immer wieder, dass es in den Unternehmen noch große Diskrepanzen zwischen den Fach- und IT-/BI-Abteilungen gibt. Häufig können die von der Fachabteilung vorgegebenen Anforderungen und Fragestellungen von der IT oder den Business-Intelligence-Bereichen, je nachdem, wer für Datenerhebung, Datenbereitstellung, Datenanalyse im Unternehmen zuständig ist, nicht schnell genug umgesetzt werden. Die Gründe dafür sind unterschiedlich: Die bestehenden BI-Umgebungen sind zu komplex, es bestehen technische Prozessrestriktionen oder Release-Prozesse sind zu langwierig. Die Folgen sind ein langsamer oder später Zugang zu entscheidenden Informationen. Dies führt nicht nur zu einer großen Unzufriedenheit der Fachanwender, sondern kann erhebliche Auswirkungen auf das Kundenmanagement und damit auf die Wettbewerbsfähigkeit des Unternehmens haben.

Die Nachfrage nach Alternativen ist daher groß und wächst beständig. Vor allem in den IT- und BI-Abteilungen, die täglich vor der Herausforderung stehen, neue Technologien einzuführen, zu unterstützen und zu betreuen, die Datenbereitstellung sicherzustellen und gleichzeitig strengen Service-Levels gerecht zu werden, ist der Bedarf an innovativen Lösungen für zufriedenere Fachbereiche so groß wie nie zuvor.

Eine Lösung kann hier der Einsatz von Self-Service BI sein, die die traditionelle BI um viele Vorteile ergänzen kann.

Abgrenzung zur traditionellen BI

Bei Self-Service BI werden parallel zu einem zentralen Data Warehouse (DWH) fachbereichsgetriebene Analysen ohne direkte Integration in die DWH-Architektur und -Prozesse bereitgestellt. Traditionelle BI ist in der Regel geprägt von historisch gewachsenen, zentralen Strukturen mit standardisierten, statischen Reports und vordefinierten Analysepfaden. Im Gegensatz dazu können die Nutzer von Self-Service BI die Datenanalyse und -auswertung, die Berichterstellung und die Integration unterschiedlicher Daten eigenständig modifizieren und flexibel an die speziellen Anforderungen des eigenen Fachbereichs anpassen. Waren traditionelle

Systeme bisher meist komplex aufgebaut und dadurch oft sehr kosten- und ressourcenintensiv, verschlanken sich bei Self-Service BI die Strukturen und verändern den Wartungs- und Betreuungsaufwand durch die IT- oder BI-Abteilung.

Insgesamt haben sich der Markt und das Geschäft stark beschleunigt. Für das Kundenmanagement bedeutet das im Konkreten schnelle Entscheidungen und zeitnahe Reaktionen. So sollte ein Telekommunikationsanbieter beispielsweise bei einem kurzfristigen Netzausfall in einer bestimmten Region schnell mit den betroffenen Kunden Kontakt aufnehmen können. Bietet das Unternehmen denjenigen Kunden, die durch den Ausfall konkrete Gesprächsunterbrechungen hatten, dann noch am gleichen Tag per SMS eine Gutschrift, können dadurch sogar noch positive Eindrücke erzeugt werden. Voraussetzung ist, dass der Fachbereich – in diesem Fall der Kundenservice – schnell auf die relevanten Informationen zugreifen kann. Jedoch können solche Situationen nicht immer vorhergesehen und Abfragen passend vorbereitet werden. In vielen Fällen lassen sich bestehende BI-Systeme fast nicht oder nur sehr aufwändig an veränderte Geschäftsprozesse oder Ad-hoc-Ereignisse anpassen. In jedem Fall ist die Fachabteilung auf die Umsetzung durch die IT- beziehungsweise BI-Abteilung angewiesen und unterliegt damit den zeitlichen und fachlichen Kapazitäten einer oftmals voll ausgelasteten Abteilung. Im Beispiel des Netzausfalls würde eine Reaktion nach mehreren Tagen oder Wochen nicht mehr die gewünschten Effekte erzielen.

Durch die Möglichkeit, einzelne Komponenten bei Self-Service BI selbst anzupassen, können die Fachabteilungen freier bei der Gestaltung und Modifikation von Abfragen agieren. Damit werden die Berichte nicht nur genauer auf ihre speziellen Fragestellungen zugeschnitten und fachlich qualitativ hochwertiger, sondern es können insgesamt schneller neue Erkenntnisse gewonnen werden. Dies bedeutet einen schnelleren Weg von der Entscheidung bis zur Umsetzung und somit geringere Aufwände und Kosten für Anpassungen und Innovationen. Gleichzeitig werden die Experten der IT- oder BI-Abteilungen deutlich entlastet und können sich somit stärker auf ihre Kernkompetenzen fokussieren.

Worauf zu achten ist

Prinzipiell bedeutet die Entscheidung für die Nutzung von Self-Service BI nicht zwingend den Erwerb eines neuen oder speziellen BI-Werkzeugs. Häufig werden für die existierenden BI-Plattformen bereits Self-Service-Funktionalitäten oder Erweiterungen angeboten, die in die bestehenden Strukturen integriert werden können. Weiterhin ist es aber möglich, speziell auf Self-Service BI zugeschnittene Werkzeuge zu nutzen und den Usern somit den bestmöglichen Service zu bieten.

Die Grundlage für eine erfolgreiche Umsetzung ist die richtige und gesteuerte Vorgehensweise. Je nach BI-Strategie eines Unternehmens wird im Vorfeld genau festgelegt, wie die bisherigen Aufgabenfelder zwischen Fachbereichen und der IT-/BI-Abteilung neu aufgeteilt werden. Eine eigene beziehungsweise ergänzende Governance ist unumgänglich. Konkret sollte die Governance in den Bereichen „Daten“, „Templates“ und „Kommunikation“ aufgestellt werden.

Um eine hohe Datenqualität zu gewährleisten, ist es erforderlich, einheitliche Kennzahlen und Dimensionen zu vereinbaren, fachbezogene Namensgebungen einzusetzen und besonders empfindliche Daten zu kennzeichnen und abzugrenzen. Gleichzeitig sollte die Governance gleiche Rahmenbedingungen für die Verwaltung und Änderungen der Daten schaffen und dadurch insbesondere Datenschutz bzw. -sicherheit garantieren. Unterstützend ist dabei der Einsatz von Templates in Form von Checklisten, die den Fachabteilungen zur Verfügung gestellt werden, sowie die Aufstellung von Standards für Reports (intern/extern) und Vorgaben für Mindestdokumentationen.

Entscheidend ist, die Vorgaben zu kommunizieren. Dazu müssen Rollen und Verantwortlichkeiten klar definiert sein und ein Austausch zwischen IT und Fachbereich bei Änderungen und Neuerungen gewährleistet werden. Verzichtet man auf eine eigene Governance, kann das zu unterschiedlichen Daten, Analyse-Fehlern und zu sinkender Datenqualität führen.

Vier verschiedene Ansätze

Derzeit sind hauptsächlich vier unterschiedliche Ansätze für die Einführung von Self-Service BI gängig. Dazu gehören die technologische Bereitstellung eines individuellen, isolierten Datenbank-Bereichs (Sandboxing), die zentrale Bereitstellung der Reports etwa in zentralen Datenablagen mit Zugriffskontrolle (Managed BI Services), die Bereitstellung von konfigurierbaren beziehungsweise eingeschränkten ETL-Applikationen zur Implementierung von dynamischen ETL-Prozessen sowie die Bereitstellung generischer Data Marts.

Bei der Auswahl der Nutzer für Self-Service BI empfiehlt es sich, auf deren technische und fachliche Skills zu achten, damit diese die Funktionen und Werkzeuge optimal nutzen können und auch Spaß daran haben, mit den Werkzeugen zu arbeiten. Prinzipiell werden keine speziellen Kenntnisse vorausgesetzt, doch sollte der User zumindest den Umgang mit Standard-Software gewohnt sein. Zusätzliche persönliche Eigenschaften wie analytisches Denkvermögen, Kreativität, Eigeninitiative und ein ausgewogenes Verhältnis zwischen Übereifer und respektvoller Herangehensweise qualifizieren den Nutzer für Self-Service BI. Auch das Verständnis gegenüber Geschäftsprozessen, der Transformation von Prozessen in Daten sowie der Analyse-Ziele (Darstellungstechniken) erhöhen die Akzeptanz des Users für diese Vorgehensweisen. Zudem ist der Einsatz von sogenannten „Key-Usern“ ratsam. Diese besitzen erweiterte Tool-Kenntnisse, kennen den Datenhintergrund, sorgen für die Einhaltung von Standards und sind Ansprechpartner sowohl für die Endnutzer als auch die BI- oder IT-Abteilung.

Vom Umsetzungs-Dienstleister zum agilen BI-Self-Service-Dienstleister

Die BI-/IT-Abteilung sollte sich beim Einsatz von Self-Service BI von einem

BI-Umsetzungsdienstleister zum agilen BI-Self-Service-Dienstleister wandeln. Dies zeichnet sich vor allem dadurch aus, dass sie administrative Aufgaben konzentrierter bearbeitet, mehr im Hintergrund agiert und die Fachanwender bei allen möglichen und auch unmöglichen Vorgehensweisen unterstützt. Dabei entsteht eine Verlagerung der Aufgaben auf die Organisation und Überwachung von Zugängen, Sicherheit und Qualität der bereitgestellten Daten (beispielsweise bei Sandboxing), aber auch auf Schulungen der Key-User und den Support der BI-Tools.

Wenn die Auswahl doch auf ein neues Tool fällt, so sollte dieses Self-Service-BI-Tool mit Hinblick auf gute Usability, Wartbarkeit, Second-Level-Support und Dokumentation getroffen werden. Es sollte sich ideal in die Tool-Landschaft des Unternehmens integrieren lassen und selbst bei hoher Auslastung eine optimale Performance gewährleisten. Dies vor allem, wenn neben den Daten aus dem DWH auch externe Datenquellen angebunden werden. Zusätzliche Funktionen und Features (etwa erweiterbare Reporting-Vorlagen oder personalisierte Dashboards) sollten den nicht-technisch versierten User nicht überfordern, sind aber besonders nützlich, wenn sie den unterschiedlichen Berechtigungsrollen zugeordnet werden können.

Chancen und Risiken

Für Unternehmen ergeben sich mit Self-Service BI viele Chancen. Zu den wichtigsten gehört die Erhöhung der Agilität und Flexibilität des BI-Systems, die es Unternehmen ermöglicht, schneller und gezielter Marktveränderungen zu erkennen und darauf eingehen zu können. Insgesamt werden sowohl die Fachabteilungen als auch das IT- oder BI-Team entlastet. Sie können sich so stärker auf ihre Kernthemen konzentrieren. Aufwände und Kosten können reduziert werden, während sich die fachliche Qualität der Reports und Entscheidungsvorlagen entscheidend verbessert. Das führt erfahrungsgemäß auch zu einer höheren Akzeptanz und Motivation der Nutzer.

Allerdings bringt die neue Unabhängigkeit auch Risiken mit sich. Ist sich ein Unternehmen dieser bewusst, so kann es sie entweder minimieren oder sogar vollständig ausschließen. So besteht zum Beispiel bei nicht ausreichender oder schlecht formulierter und kommunizierter Governance die Gefahr uneinheitlicher Kennzahlen und inkonsistenter Berichte sowie eines Auftretens von ungesteuerten Insellösungen. Klare Vorgaben und Regeln können dies leicht verhindern und führen gleichzeitig zu einem sinnvollen Wartungsaufwand in den IT- oder BI-Abteilungen. Sind die Anwender nicht ausreichend geschult, kann es aufgrund der neuen Verantwortung auch zu einer Überforderung oder Überlastung kommen. Die Praxis hat gezeigt, dass gezielte Einweisungen, Coaching und Workshops ein gutes Ergebnis erzielen. Besonders wichtig ist es, darauf zu achten, die technische und inhaltliche Qualität der Lösung vor allem zu Beginn regelmäßig zu prüfen und durch eine konsistente Dokumentation zu ergänzen, um isoliertes Spezialwissen zu vermeiden.

Ergänzung der Prozesse

Die Ergänzung der bestehenden Anforderungs- und Incident-Prozesse ist ein weiterer Faktor zur Umsetzung von nutzerfreundlicheren BI-Vorgängen. Der Anforderungs-Prozess sollte so verändert werden, dass die Aufnahme von Self-Service-BI-Reports möglich ist und nicht als problematisch abgewiesen wird. Dazu sind schnelle Antworten und schnellere Einarbeitung in unbekannte Themenfelder für BI-Analysten erforderlich. Benötigt wird dazu allerdings eine Grunddokumentation der SSBI-Reports.

In den Incident-Prozessen ist es erforderlich, Personen auszubilden, die Incidents von nicht Fachanwendern entgegennehmen – auch wenn die referenzierten Kennzahlen und Berichte nicht in den BI-/IT-Abteilungen bekannt sind. Dies fördert wiederum die Flexibilität durch schnelle Antworten, eine schnelle Einarbeitung in unbekannte Themen sowie die Übergabe von Informationen in den Anforderungs-Prozess. Eine professionelle Anpassung der Prozesse bringt gewünschten Erfolg und Akzeptanz der Nutzer.

Fazit

In vielen Fällen nutzen Fachbereiche heute schon eigene Auswertungen und eigene Berichtswesen – meist aber ungesteuert und unabhängig von den Prozessen und Möglichkeiten vorhandener BI- und IT-Abteilungen. Eine gesteuerte Einführung oder Ergänzung um Self-Service BI ist empfehlenswert. Die Zufriedenheit der Fachbereiche erhöht sich und die BI- und IT-Abteilungen können ihre Dienstleistungen noch gezielter anbieten.

Self-Service BI bietet den Anwendern eine Unabhängigkeit und Selbstständigkeit bei der Nutzung von Unternehmens-Informationen und der Analyse von Daten, wie beispielsweise bei der Auswertung von Kundendaten oder der Erstellung von Berichten. Durch eine größere Flexibilität der BI-Strukturen wird den Fachabteilungen ein leichterer Zugang zu wichtigen Informationen über vorhandene DWH-Daten hinaus mit zusätzlichen Quelldaten ermöglicht. Das Ergebnis ist, schneller relevante Entscheidungen treffen zu können. Gerade im Kundenmanagement ist dies ein wichtiges Differenzierungsmerkmal, denn hier treten Veränderungen sehr schnell und teilweise in kurzen Zyklen auf.

Mit dem Einsatz von Self-Service BI können Unternehmen schneller auf die Marktgegebenheiten und das Verhalten ihrer Kunden reagieren und damit einen großen Wertbeitrag zum Erfolg des Unternehmens leisten.

Matthias Spieß
matthias.spiess@shs-viveon.com

Der Oracle DBA

Gelesen von Thomas Tretter

Alle paar Jahre kann man sich auch als erfahrener DBA mal wieder ein neues Handbuch kaufen. Das war mein erster Gedanke, als ich dieses Buch gesehen habe. Auffallend ist sofort die lange Liste der Autoren. Viele von ihnen sind mir persönlich oder aus verschiedenen Vorträgen bekannt und konnten ihr Praxiswissen schon häufiger unter Beweis stellen. Am Ende des Buches werden die Autoren auch kurz vorgestellt, außerdem ist dort nachzulesen, wer welche Kapitel beigesteuert hat.

Misstrauisch bin ich allgemein bei Büchern von mehreren Autoren bezüglich des durchgehenden Schreibstils, des Aufbaus und speziell auch in Bezug darauf, ob die gegenseitigen Verweise stimmen. Um es gleich vorwegzunehmen: Ich bin positiv überrascht! Das Buch ist aus einem Guss. Jedes Kapitel beginnt mit einem einführenden Überblick und endet mit einem Resümee.

Mein zweiter Blick gilt gewöhnlich dem Inhaltsverzeichnis. Welche Themen werden grundsätzlich behandelt? Erscheinen mir der Aufbau und die Reihenfolge schlüssig? Auch hier scheint nichts zu fehlen, was bei einer Seitenanzahl von rund 800 Seiten auch zu erwarten ist.

Ich habe mir dann im Laufe der Zeit einige Kapitel komplett durchgelesen, um Verständlichkeit und Aufbau exemplarisch zu erfassen. Das erste Kapitel „Schnelleinstieg“ umfasst neben einer grundsätzlichen Einführung die praktische Erstellung einer Test-Datenbank. Die Erklärung und Befehle sind sowohl für Windows als auch für das Linux-Betriebssystem vorhanden. Dies zieht sich übrigens nahezu durchgängig durch das Buch. So findet jeder sofort ein leicht verständliches Beispiel aus seiner bevorzugten Betriebssystem-Welt.

16 Kapitel behandeln alle Aspekte und Möglichkeiten der Version 11*g* R2. Anhand des Inhaltverzeichnisses kann man sich gut orientieren und dann zielgenau das Kapitel mit den Themen durchlesen. Ich möchte an dieser Stelle einige Kapitel hervorheben, die mich zum Zeitpunkt des Lesens aktuell interessiert haben. Zunächst habe ich „Architektur und Administration“ gelesen, da ich es für eine notwendige Grundlage der Datenbank-Administration halte. Aufbau und speziell die Erklärungen hierzu sind gut gelungen. Auch wenn ich selbst schon einige Jahre Erfahrung habe, konnte ich hier noch einiges Wissenswertes erfahren. Allein das „Drüber-Nachdenken“ über eigentlich bekannte Sachverhalte ist jedem Leser empfohlen.

Aus konkretem Anlass im aktuellen Projekt habe ich die Kapitel „Optimierung“ und „Monitoring“ durchgearbeitet. Meine Erwartungshaltung hinsichtlich Optimierung wurde zunächst nicht befriedigt, ich hatte wahrscheinlich mehr konkrete „Kochrezepte“ erwartet. Letztendlich hat mich aber der Aufbau des Kapitels doch überzeugt, da die Vorgehensweise bei der Optimierung sehr anschaulich beschrieben ist. Positiv ist mir auch die Interpretation von Reports (tkprof, WAR) aufgefallen. Was ist das Ziel, wie komme ich dort hin. Es fehlt nicht an konkreten Hinweisen, sowohl für Anhänger des Enterprise Managers als auch für Fans von SQL-Skripten. Wer dann allerdings tiefer in das Thema einsteigen will, wird um das Studium weiterer Literatur nicht herumkommen.

In den Kapiteln „Troubleshooting“ und „Monitoring“ werden naturgemäß recht viele Enterprise-Manager-Funktionalitäten besprochen, wobei an manchen Stellen auch auf die Möglichkeit der Script-basierten Aufrufe hingewiesen ist. Aufgrund der Lizenzpolitik von Oracle bezüglich der zusätzlichen Packs (Enterprise Edition + Tuning/Diagnostic Pack) wird hier wohl mancher Leser außen vor bleiben. Es wird jedoch bei der Erklärung der Funktionalitäten stets auf dieses Problem hingewiesen. Die erwähnten weiterführenden Skripte sind alle herunterladbar.

Fazit

Das Buch umfasst das gesamte Spektrum der Oracle-11*g*-R2-Funktionalität, außerdem werden Anfänger und Fortgeschrittene gleichermaßen angesprochen. Dieser Spagat stellt natürlich eine Herausforderung dar. Ich halte das Buch trotzdem für gelungen: Niemand setzt alle Funktionalitäten in vollem Umfang ein. Bei vielen Einzelthemen hat es mich gekribbelt und ich wollte diese am liebsten gleich ausprobieren. Zudem liest sich ein deutschsprachiges Buch neben der ganzen englischen Dokumentation doch flüssiger, speziell bei den weiterführenden Erklärungen.

Thomas Tretter
thomas.tretter@doag.org

Titel:	Der Oracle DBA
Autoren:	Andrea Held, Mirko Hotzy, Lutz Fröhlich, Marek Adar, Christian Antognini, Konrad Häfeli, Daniel Steiger, Sven Vetter, Peter Welker
Umfang:	802 Seiten
Sprache:	Deutsch
Preis:	69 Euro (auch als eBook verfügbar)
ISBN:	978-3-446-42081-6

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.

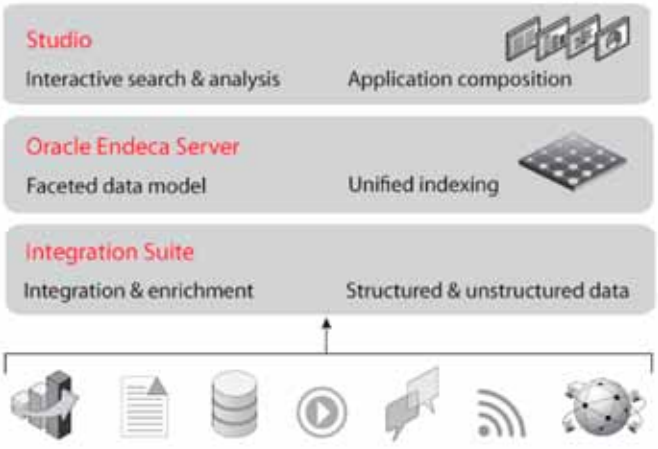


Abbildung 1: Komponenten von Oracle Endeca Information Discovery

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

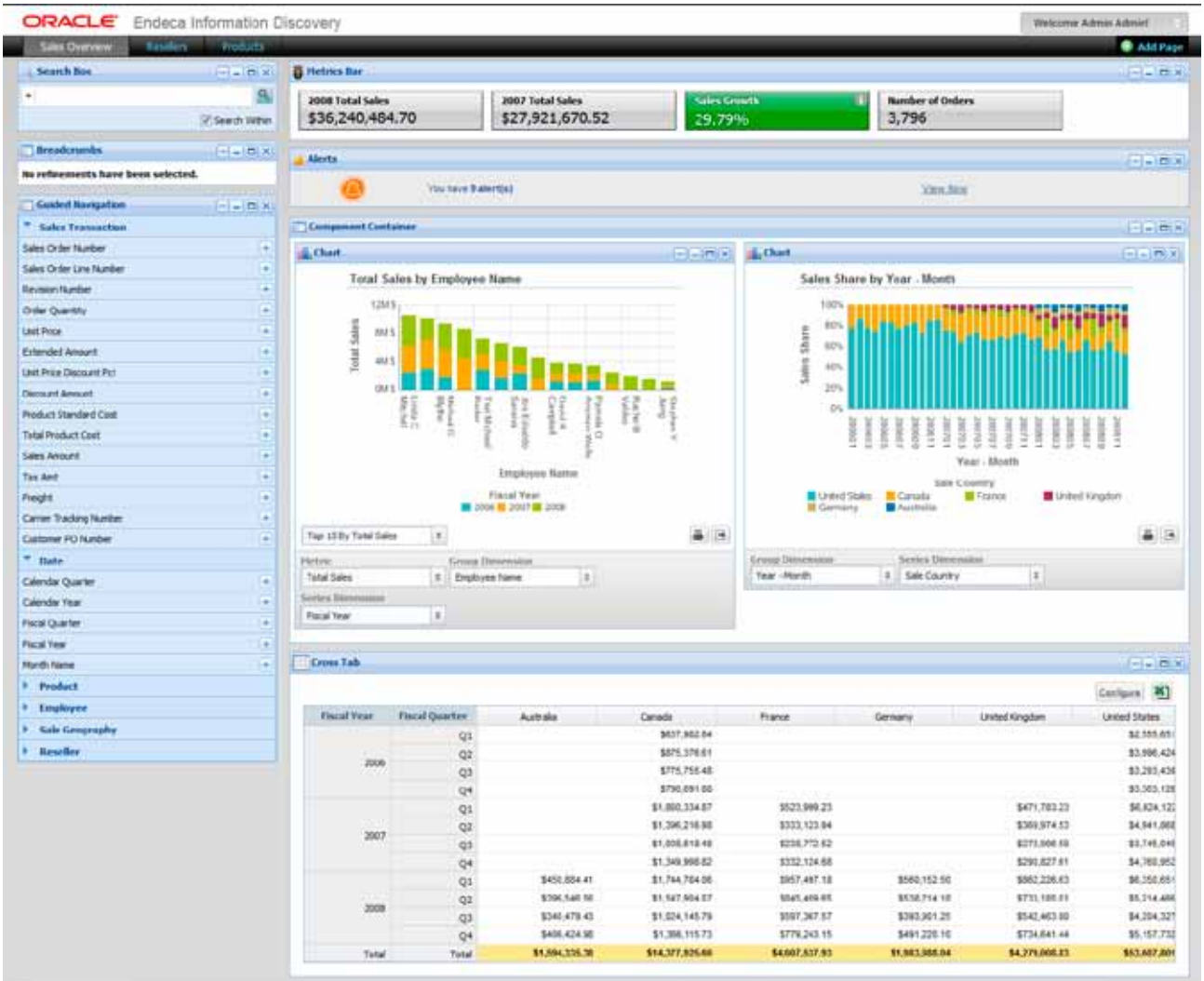


Abbildung 3: Analyse von Social-Media-Daten

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

Oracle E-Business Suite
Oracle Business Intelligence
Oracle Custom Development
Oracle Data Base Services

apps associates
extreme expertise

APEX Webinare

Oracle bietet mit APEX ein zeitgemäßes, hochprofessionelles Tool, mit dessen Hilfe sich in kürzester Zeit webfähige Applikationen für den Einsatz in Unternehmen erstellen lassen. Doch der Teufel bei Entwicklung, Deployment und Betrieb steckt im Detail.

Apps Associates bietet Interessierten ab September 2012 regelmäßige Webinare zu speziellen Aspekten aus der APEX-Welt an. Themen werden die Betriebsoptimierung von APEX-Landschaften, strategische Einsatzmöglichkeiten, Grenzen des APEX-Einsatzes sowie Praxisbeispiele sein.

Die Reihe wendet sich an CIOs, Entwickler, Projektleiter und Interessierte aus Fachbereichen. Anmeldung und Teilnahme sind kostenlos. Bitte informieren Sie sich auf der u. a. Website.

www.appsassociates.de/apex

ORACLE Platinum Partner

Was versteht man unter Big Data, wer braucht es und wie unterscheidet es sich vom klassischen Data Warehousing & BI? Zudem stellt sich die Frage, ob große Datenmengen heutzutage nicht auch in den modernen Data-Warehouse-Appliances gut aufgehoben sind. Der Artikel beantwortet diese Fragen und betrachtet typische Problemstellungen für Hardware, Software und Datenmodellierung jeweils mit einem Blick auf die Lösungsansätze von Oracle Exadata, Teradata & Co.

Big Data (Warehouse?)

Peter Welker, Trivadis GmbH, Stuttgart

Keine IT-Veranstaltung oder BI-Publikation ohne „Big Data"! Auf der DOAG 2012 BI gab es einen eigenen Track für dieses Thema und auf der TDWI Konferenz 2012 im Juni fanden sich sieben Vorträge dazu – nicht gerechnet die Präsentationen, die dieses Thema zumindest streiften. Nach Gartner ist der Hype gerade erst angelaufen. Aber was ist eigentlich die Ursache für diese Welle? Das Datenvolumen in IT-Systemen steigt doch schon seit Jahrzehnten kontinuierlich an.

Geschätzte 1,8 Zettabyte (entspricht 1.800 Exabyte) an Daten wurden im Jahr 2011 weltweit erzeugt und repliziert – dazu gehört auch die Verbreitung im Internet. Das ist neunmal so viel wie noch vor fünf Jahren [1]. Würde man die für die Speicherung dieser Daten benötigten Server-Festplatten hochkant nebeneinander stellen, reichte dieser Gürtel zweimal um die Erde. Wirklich neu produziert und mehr oder weniger dauerhaft gespeichert wurden 2010 geschätzte 13 Exabyte Daten. Das McKinsey Global Institute (MGI) sieht in der Nutzung der in diesen Daten enthaltenen Informationen ein jährliches Potenzial von rund 300 Milliarden Dollar allein im amerikanischen Gesundheitswesen und 250 Milliarden Euro in Europas öffentlichem Sektor [2].

Die beeindruckenden Zahlen wecken möglicherweise die Sorge, gerade etwas Wesentliches zu verpassen. Dieser Quantensprung an Quantität darf aber nicht darüber hinwegtäuschen, dass auch heute nur ein winziger Bruchteil dieser Informationen für die Entscheidungsfindung in modernen Unternehmen relevant sind.

Un-/Poly-/Semi-/Irgendwie strukturierte Daten

Ein immer größerer Anteil dieser neuen Daten basiert aus BI-Sicht auf weitgehend unstrukturierten, aber immens umfangreichen Formaten wie Bildern, Audios, Videos oder Layouts. Auf YouTube wurde im Frühjahr 2010 jede Minute vierundzwanzig Stunden Videomaterial hochgeladen, heute sind es mehr als fünfzig. Daneben steigt auch der Platzbedarf innerhalb desselben Mediums drastisch. Ein Full-HD-Video mit zeitgemäßer Tonspur benötigt mit bis zu 6,75 Megabyte pro Sekunde leicht dreißigmal so viel Speicherplatz wie derselbe Film im früheren PAL-Fernsehen, transportiert aber für den überwiegenden Teil aller denkbaren Anwendungsfälle immer noch dieselben Informationen. Und mit „4K2K" und „8K4K" stehen bereits die Nachfolger mit vier- beziehungsweise sechzehnfacher Auflösung in den Startlöchern. Abgesehen von wenigen strukturierten Anteilen (Tags) sind solche Daten für die Unternehmenssteuerung eher irrelevant. Das mag sich gegebenenfalls durch Anwendungen wie „automatisierte Gesichtserkennung" in Zukunft partiell ändern, wird sich aber selbst dann wohl meist in der Aggregation dieser Daten in kleinen, klar strukturierten Einheiten niederschlagen.

Interessant ist das Potenzial für uneinheitlich strukturierte Daten wie den personifizierbaren, durch Volltext und Multimedia angereicherten „Social-Media Content" aus Facebook, Twitter & Co. Dieser ist zwar mit klassischen BI-Ansätzen (noch) nicht leicht zu analysieren, bietet aber bereits heute die Möglichkeit, Meinungsführer und Trends zu identifizieren und wird auch schon vereinzelt im Marketing genutzt. Hier sind zur Analyse von Beziehungen sogenannte „Graphen-Datenbanken" nützlich.

Besonderes Augenmerk benötigen auch die Informationen, die sich inzwischen mehr und mehr durch permanent von Maschinen produzierte oder produzierbare Daten ergeben. Man spricht auch medienwirksam vom „Internet of Things". Dazu darf man sowohl geografisch-persönliche Ereignisse wie die Lokalisierung von Smartphones und das Lesen von RFIDs, aber auch die heute fast schon klassischen Web-Logs oder die immer mitteilungssüchtigere Produktions-Sensorik zählen.

Strukturierte Daten

Natürlich spielen auch herkömmliche Daten eine bedeutende Rolle im Mengenwachstum. Unternehmen und Behörden erzeugen heute deutlich größere Informations-Einheiten, speichern also immer mehr Informationen zu jedem Kunden und mehr Details zu jeder Transaktion. Gleichzeitig wachsen auch die Anzahl der Informations-Einheiten und damit die Granularität in Stammdaten und Bewegungsdaten gleichermaßen. Diese Daten bilden nach wie vor den Kern der traditionellen BI – und sind nicht selten über bestimmte Zeiträume hinweg zur Gewährleistung höchster Transparenz erforderlich, wie zum Beispiel Diagnose- und Leistungsdaten im Gesundheitswesen, Call Data Records bei den Telekommunikations-Dienstleistern oder Kontenbewegungen im Finanz-

wesen. Gerade durch das Zusammenführen dieser Informationen zu einem einheitlichen Kundenverständnis entstehen wertvolle zusätzliche Erkenntnisse.

Das Dilemma

Es stellt sich die Frage, wie mit dem Dilemma des Datensammlers umgegangen werden kann: „Müssen alle Daten über lange Zeit gesammelt werden, um schon heute für zukünftige Anforderungen gerüstet zu sein? Oder genügt es, diese neuen Datenquellen erst anzuzapfen, sobald konkrete Anforderungen aufkommen?“ Letzteres erlaubt deutlich geringere Investitionen, allerdings auf Kosten des Risikos, von neuen Erkenntnissen nicht schnell genug profitieren zu können und auf deren Geschichte zu verzichten.

Big Data, eine disruptive Technologie?

Die große Riege der NoSQL-, BigTable-Clone- und sonstigen verteilten Datenbanken – also Hadoop, Dynamo, Cassandra & Co. – haben ihre Einsatzfähigkeit für (uneinheitlich) strukturierte Daten in Einzelfällen längst eindrucksvoll unter Beweis gestellt. So betreibt beispielsweise Yahoo Hadoop-Cluster mit mehreren Hundert Petabyte Daten.

Es stellt sich demnach die Frage nach der richtigen Technologie: „Bin ich mit meinen gegenwärtigen BI-Lösungen auf dem Holzweg? Wird „Big Data“ die Art und Weise, wie wir mit unseren entscheidungsrelevanten Daten umgehen, von Grund auf verändern?“

Heute deutet alles darauf hin, dass die gängigen Big-Data-Ansätze mit Hadoop & Co. mittelfristig nicht geeignet sind, die Aufgaben klassischer BI-Lösungen zu übernehmen. Warum? Die Nutzung von Hadoop mittels Map-Reduce-Framework ist zwar prinzipiell extrem flexibel – schließlich kann man alles tun, was mit Programmierung und Dateien umsetzbar ist –, dadurch aber auch mit sehr hohem Entwicklungsaufwand verbunden und direkt für Endanwender praktisch nicht nutzbar. Alternative Datenzugänge wie Hive, das eine SQL-Schnittstelle für Hadoop darstellt und Map-Reduce-Zugriffe generiert, sind hingegen weit von der Funktionalität klassischer Datenbank-Systeme entfernt.

Letztlich ist der Anspruch der BI an Antwortzeiten (etwa Sub-Second bei MOLAP-Abfragen), Variabilität, Einfachheit für Entwickler und Anwender sowie die Ausrichtung auf seit fast zwanzig Jahren etablierte Methoden wie Pivotisierung und dimensionale Modelle im Zusammenspiel mit relationaler und multidimensionaler OLAP-Technologie sehr ausgeprägt. Langfristig könnten die BI-Anwenderwerkzeuge Datenquellen wie Hadoop & Co. als Alternativen einbeziehen. Solange lassen sich mit Big-Data-Ansätzen aber eher gänzlich neue Aufgabenfelder erschließen oder zweckentfremdende Prozesse aus bestehenden Data Warehouses extrahieren. Besonders kann jedoch die Vor- und Aufbereitung entscheidungsrelevanter Informationen aus bisher unzugänglichen Daten den etablierten BI-Systemen schon heute gute Dienste leisten. Die neuen Technologien sind also nicht disruptiv, sondern vielmehr ergänzend und als Quelle innerhalb der Business Intelligence nutzbar.

Lösungsansatz „Smart Data“

Ein naheliegender, praktikabler Kompromiss stellt diese Strategie dar. Dabei können die hohen Anforderungen an die Analyse strukturierter Daten heute mit DWH-Appliances bis in den zwei- und dreistelligen Terabyte-Bereich hinein bewältigt werden – allerdings zu entsprechenden Kosten. Die große Masse der uneinheitlich strukturierten Daten wird dazu mit den jeweiligen Lösungen aus dem „NoSQL & BigTable“-Umfeld (wie Hadoop) verwaltet, selektiert und in bestimmtem Umfang auch analysiert. Bei Bedarf können die dort gehaltenen Daten dann aufbereitet und ins klassische DWH geladen werden. Aus Big Data würde somit für BI-relevante Anwendungsfälle „Smart Data“ (siehe Abbildung 1).

BigData und DWH-Appliances

Inzwischen stellen einige Anbieter zwei verschiedene Varianten von Appliances zur Auswahl: Big-Data-Lösungen und „Massive Parallel Processing (MPP)“-Data-Warehouse-Cluster-Lösungen. Die für „Big Data“ deklarierten Lösungen wie Oracle Big Data Appliance oder EMC Greenplum HD bauen weitgehend auf Hadoop auf und stellen dafür spezielle Hadoop-Distributionen mit zahlreichen Erweiterungen und eigenen Implementierungen beispielsweise von Key-Value-Stores (Teradata Aster) zur Verfügung.

Im zweiten Teil des Artikels konzentrieren wir uns auf diese Data-Warehouse-Appliances und betrachten, welche Lösungen hier angeboten werden und welche Architekturen dafür zum Einsatz kommen. Typische Ver-

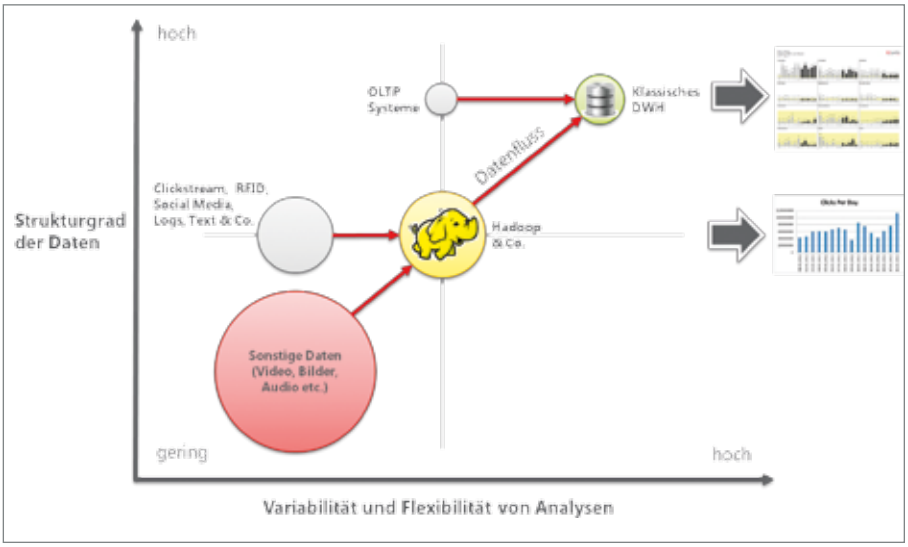


Abbildung 1: Mehrstufiges Konzept zur Analyse von Daten unterschiedlichen Strukturierungsgrades

treter dieses Genres sind Oracle Exadata, Teradata, IBM Netezza, HP Vertica, EMC Greenplum, Microsoft PDW oder Exasol.

Diese Data-Warehouse-Appliances fokussieren sich auf klassische DWH-Methoden und nutzen dafür spezielle RDBMS-Implementierungen mit weitgehend standardisierter SQL-Schnittstelle. Das Besondere daran: Es handelt sich um Datenbank-Cluster, die nach dem Scale-Out-Prinzip große Datenmengen auf einzelne, relativ kleine Server-Knoten verteilen. Durch geschickte Partitionierung der Aufgaben werden die Daten dann möglichst direkt auf den einzelnen Knoten vorverarbeitet, also beispielsweise gefiltert, per Join miteinander verknüpft und aggregiert. Abschließend werden die stark verkleinerten Resultate an einer Stelle zusammengeführt und dem Anwender zurückgegeben. Dafür müssen nur relativ kleine Ergebnismengen über das interne Cluster-Netzwerk geschickt werden.

Generelle Aufgabenstellungen bei großen Datenmengen

Prinzipiell sind die Aufgabenstellungen an Skalierung durch Verteilung bei allen Cluster-Lösungen, ganz gleich ob HDFS oder Datenbank-Cluster, sehr ähnlich. Es geht dabei vereinfacht gesagt um Folgendes:

- Eine einzelne Aufgabe auf vorhandene Ressourcen zu verteilen (Partitionierung)
- Diese Verteilung möglichst gleichmäßig durchzuführen (Skewing vermeiden)
- Auch bei mehrstufigen Aufgaben verfügbare Ressourcen immer optimal auszulasten (Pipelining)
- Dabei möglichst viele Operationen innerhalb eines Knotens auszuführen (Lokalität)

Dabei gibt es natürlich jede Menge Möglichkeiten, die Auslastung der Ressourcen gering zu halten beziehungsweise optimal aufeinander abzustimmen, um gute Antwortzeiten zu erreichen wie:

- Organisation (Spalte/Zeile) und Kompression der Daten, um einen höheren De-Facto-Durchsatz zu erzielen
- Maßnahmen zur Reduktion der zu verarbeitenden Daten wie Indexierung, horizontale und vertikale Partitionierung, Vorhalten von Min-/Max-Werten für Datenbereiche etc.
- Priorisierung bestimmter Abfragen nach Anwendergruppe oder zu erwartender Laufzeit (Ressourcen-Management) etc.

Da analytische Aufgaben meist eine ganze Reihe von elementaren Operationen beinhalten (Scans, Joins, Filter, Gruppierungen, Sortierungen), kommt es auch und besonders darauf an, wie der Query-Optimizer eine Aufgabe zerlegt, wie er also die einzelnen Schritte zur Erfüllung der Aufgabe auswählt und kombiniert. Soll er beim Join die Daten einer Tabelle an alle Knoten schicken oder nur Teile davon an bestimmte Knoten? Soll er zuerst die Tabelle A oder die Tabelle B scannen? Um diese und noch viel mehr Fragen beantworten zu können, muss bekannt sein:

- Nach welchen Kriterien die Daten auf den Knoten verteilt werden (Distribution)
- Welche Tabelle welche Datenmengen hat und welche Daten sich in den einzelnen Spalten befinden (Statistiken)

Beispiele für Lösungsansätze

Generell sind die Anbieter mit ihren Lösungen auf die Anforderungen eingestellt. Alle nutzen umfangreiche Statistiken über die Datenverteilung für die Ermittlung von Ausführungsplänen und bieten Lösungen zur Reduktion des Datenverkehrs über den Interconnect. Zur Verteilung/Distribution der Daten auf die einzelnen Knoten teilen sich die Lösungen in zwei Lager: Die einen nutzen eine klassische Shared-Nothing-Architektur, haben also pro Knoten lokale Platten und verteilen die Daten einer Tabelle zufällig oder nach dem Hashwert von definierbaren Spalten möglichst gleichmäßig auf die Knoten. Die anderen – eigentlich nur Oracle Exadata – fahren einen gemischten Ansatz. Dabei kommen zwei Typen von Knoten zum Einsatz: Die einen (klassische Datenbank-Knoten, meist als RAC konfiguriert) sehen die anderen (Storage Server) als eine Art „Shared Disks“. Auf den Storage-Servern wiederum werden die Daten dann jedoch gleichmäßig verteilt. Dieser eher hybride Ansatz erlaubt Oracle die „1:1“-Nutzung seiner Standard-Funktionalität. Natürlich speichern alle Anbieter standardmäßig alle Daten redundant auf mehreren Knoten, um Ausfallsicherheit zu gewährleisten.

Alle erlauben auch die Kompression von Daten. Die Methoden dafür sind allerdings sehr unterschiedlich. Oracle Exadata, EMC Greenplum und mittlerweile auch Teradata bieten beispielsweise verschiedene Varianten (zeilenorientiert, spaltenorientiert) in verschiedenen Stärken an. Die rein spaltenorientierten Lösungen wie HP Vertica und Exasol komprimieren natürlich rein spaltenorientiert, was je nach Datenlage sehr effizient sein kann, und tun dies „out of the box“. IBM Netezza hingegen nutzt spezielle Hardware (FPGAs), um beispielsweise die Kompression auf Disk möglichst effektiv zu gestalten.

Beim Ressourcen-Management können nur wenige Anbieter direkt Einfluss auf die einzelnen Elemente einer Operation nehmen, also CPU, Disk und Netzwerk prozentual auf einzelne Aufgaben verteilen, um so beispielsweise taktischen Anwendern eine garantierte Verfügbarkeit zu gewährleisten. Teradata hat hier sehr umfangreiche Möglichkeiten und Oracle Exadata bietet – im Gegensatz zur normalen Datenbank – neben der prozentualen CPU-Zuteilung auch eine prozentuale IO-Zuteilung an. Bei den meisten anderen Lösungen beschränken sich die Möglichkeiten auf die Nutzung von unterschiedlichen Queues pro Anwendergruppe oder auf vergleichbare Ansätze.

Standby, Backup & Co.

Was bedeutet die Einführung einer solchen Lösung nun aber aus systemadministrativer Sicht?

Backups kann man natürlich bei allen Systemen ganz klassisch (auch im Betrieb) ziehen. Aber ein Snapshot, wie

man ihn aus der SAN-Welt kennt, ist bei lokalen Disks nicht möglich. Eine Datenbank zu klonen bedeutet also schlicht, sie zu kopieren.

Bei den Hochverfügbarkeits-Lösungen (HA) wird es uneinheitlich. Manche erlauben, den Cluster auseinanderzuziehen (Stretch-Cluster). Das ist zwar bei relativ kleinen Abständen (etwa 100 m über eine Feuerschutzwand) noch performant möglich, bedingt aber auch eine Verdopplung der Redundanz, wenn nach dem Ausfall einer Seite die andere alleine nicht ohne Redundanz der Disks weiterlaufen soll. Wenn die Anforderungen allerdings eine Verteilung auf weit entfernte Standorte nötig machen (etwa 100 km), kann ein Stretch-Cluster auch sehr ineffizient werden, denn einerseits steigt mit der Entfernung naturgemäß die Latenz und andererseits sind optimierte Interconnects in WANs nicht mehr so ohne Weiteres realisierbar.

Oracle bietet immer die Möglichkeit, eine Standby-Datenbank aufzubauen (Data Guard). Teilweise ist bei anderen Anbietern eine ähnliche Lösung vorhanden oder der Einsatz von Replikationssoftware möglich. Schlimmstenfalls muss man jedoch alle ETL-Prozesse an zwei Standorten laufen lassen.

Kraft versus Intelligenz

Neben den ganzen brachialen Möglichkeiten der Skalierung und der optimalen Ressourcen-Nutzung darf man aber auch die „Intelligenz“ einer Lösung nicht vernachlässigen. So ist es beispielsweise schlicht Geldverschwendung, wenn man in einer typischen DWH-Lösung seine Daten ausschließlich auf dem untersten Granularitäts-Level analysiert, obwohl eine redundante Haltung vor-aggregierter Daten 80 Prozent der Abfragen mit einem Tausendstel des Aufwands bedienen könnte.

Literaturverzeichnis

[1] IDC Reportserie: „Digital Universe”, 2007 ff. Sponsored by EMC
[2] MGI Report: „Big Data. The next frontier for innovation, competition, and productivity.” 2011
[3] http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/de//archive/mapreduce-osdi04.pdf

Peter Welker
peter.welker@trivadis.com

IT-Consulting

› **Performance Tuning**
- Oracle Datenbank Tuning
- Oracle SQL + PL/SQL Tuning

› **Real Application Clusters**

› **Data Guard + Fail Safe**

› **Datenbank Management**
- Konfiguration
- Backup & Recovery
- Migration und Upgrade

› **OEM Grid Control**

› **Oracle Security**

› **Services**
- Remote DBA Services
- Telefon-/Remotesupport

Nutzen Sie unsere Kompetenz für Ihre Oracle Datenbanken.

Schulungen

› **Oracle SQL**
› **Oracle PL/SQL**
› **Oracle DBA**
› **Oracle APEX**
› **Backup & Recovery**
› **RMAN**
› **Neuerungen 10g/11g**
› **Datenbank Tuning**
› **Datenbank Monitoring**
› **Datenbank Security**

Wir bieten Ihnen öffentliche Kurse sowie Inhouse-Schulungen.

Software-Lösungen

› **Individualsoftware**
- .NET und Visual Basic
- Java

› **Oracle APEX**

› **PL/SQL**

Unser Ziel: Individuelle Softwareentwicklung mit Fokus auf Ihre Zufriedenheit.

Oracle Lizenzen

› **Oracle Datenbanken**
- Standard Edition One
- Standard Edition
- Enterprise Edition
- Personal Edition

› **Oracle Produkte**
- Enterprise Manager
- Oracle Tools

Optimale Lizenzierung durch individuelle Beratung.

MUNiQSOFT

MuniQSoft GmbH • Grünwalder Weg 13a • 82008 Unterhaching • Telefon: 089 / 6228 6789-0 • http://www.muniqsoft.de • info@muniqsoft.de

ORACLE Gold Partner

Seit einigen Monaten wird „Big Data“ intensiv, aber auch kontrovers diskutiert. Dieser Artikel zeigt nach einem einführenden Überblick anhand von Anwendungsfällen auf, wo die geschäftlichen Mehrwerte von Big-Data-Projekten liegen und wie diese neuen Erkenntnisse in die bestehenden Data-Warehouse- und Business-Intelligence-Projekte integriert werden können.

Analytische Mehrwerte von Big Data

Oliver Röniger und Harald Erb, ORACLE Deutschland B.V. & Co. KG

Der McKinsey-Report „Big Data“ betont die enorme gesellschaftliche und geschäftliche Bedeutung, die sich aus den explodierenden Datenmengen in nahezu allen Branchen ergibt [1]. Um tatsächlich von „Big Data“ zu sprechen, sind drei Merkmale zu erfüllen („3 Vs“):

- *Volume*
  Riesige Datenmengen (xx Terabyte), die sich bislang nicht für Data-Warehouse-Analysen erschließen lassen, weil deren relevante Informationsdichte einfach zu gering ist, als das sich deren Speicherung und Verarbeitung aus wirtschaftlicher Sicht lohnt.
- *Velocity*
  Die hektische zeitliche Frequenz, in der Daten in operativen Geschäfts-Prozessen entstehen. Mehrwerte werden sowohl aufgrund der sehr hohen Granularität der Daten als auch in deren umgehender Verarbeitung und Erkenntnisgewinnung in Echtzeit gesehen.
- *Variety*
  Die Vielfalt der zusätzlichen (unstrukturierten) Datenformate, die sich jenseits der üblichen wohlstrukturierten Transaktionsdaten aus Social-Media-Daten, Maschine-zu-Maschine-Kommunikationsdaten, Sensordaten, Webserver-Logdateien etc. ergeben.

Diese Daten sind inhaltlich neu, sie sind unstrukturiert, es sind unsagbar viele – die wirklich interessanten Informationen darin sind hingegen nur äußerst dünn gesät. Insofern liegt es nahe, sich an das folgende einfache Vorgehensmodell zu halten:

1. Gezieltes Sammeln der neuartigen Massendaten aus den relevanten Datenquellen
2. Filtern dieser Daten aufgrund definierter interessanter Merkmale
3. Selektive Weiterverarbeitung beziehungsweise Übernahme der interessanten Informationen in die vorhandenen internen IT-Systeme
4. Die verarbeiteten Daten aus dem 1. Schritt wegwerfen und den Prozess fortsetzen

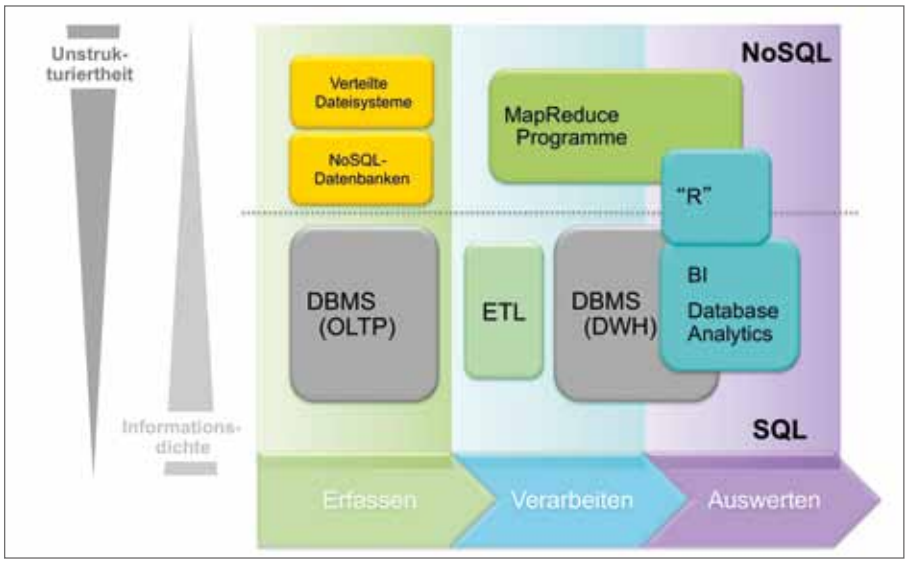


Abbildung 1: Gegenüberstellung der Komponenten

Um diese unstrukturierten, schemalosen Daten überhaupt sammeln zu können, wurden von Google und anderen Internet-Pionieren NoSQL-Datenbanken (wie Cassandra) entwickelt und mit Hadoop sowohl ein verteiltes Dateisystem (HDFS) als auch ein Entwicklungs-Framework (MapReduce) bereitgestellt (siehe Positionierung der Oracle Big Data Appliance [2]). Abbildung 1 stellt die maßgeblichen Komponenten der NoSQL- und SQL-Welt gegenüber.

Zunächst soll eine mögliche gemeinsame Architektur betrachtet werden, um diese Technologien parallel oder auch gemeinsam zu betreiben, bevor aus Anwendungssicht die Frage geklärt wird, was dieses pragmatische Vorgehensmodell konkret für verschiedene Anwendungsfälle bedeutet.

Zusammenspiel Big Data/Data Warehouse

Bei einer klassischen Konzeption eines Data-Warehouse und Business-Intelligence-Systems, leicht modifiziert nach [3], bleiben durch Big Data die bestehenden Data-Warehouse- und Business-Intelligence-Prozesse zunächst unangetastet. Die neuartigen Datenquellen erweitern aber zum einen den analyserelevante Datenraum, was Erkenntnisgewinn verspricht, zum anderen treten an die Seite von klassischen BI-

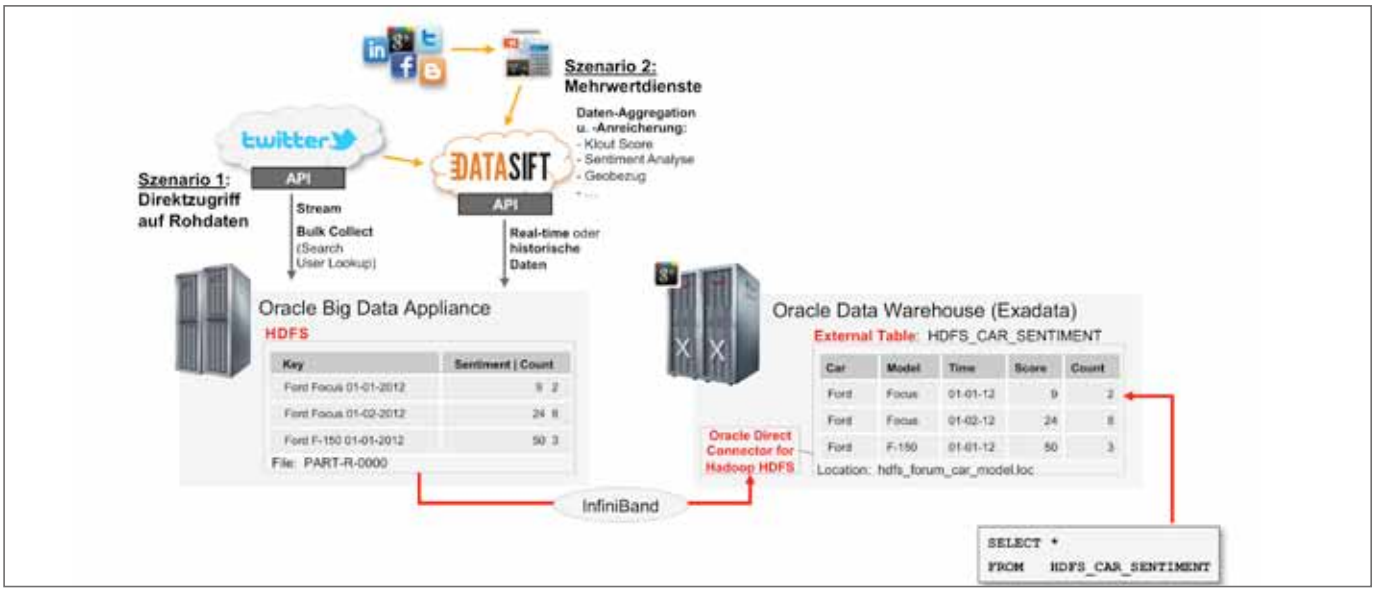


Abbildung 2: Vom Twitter-Feed zum Big-Data-Zugriff via External Table im Data Warehouse

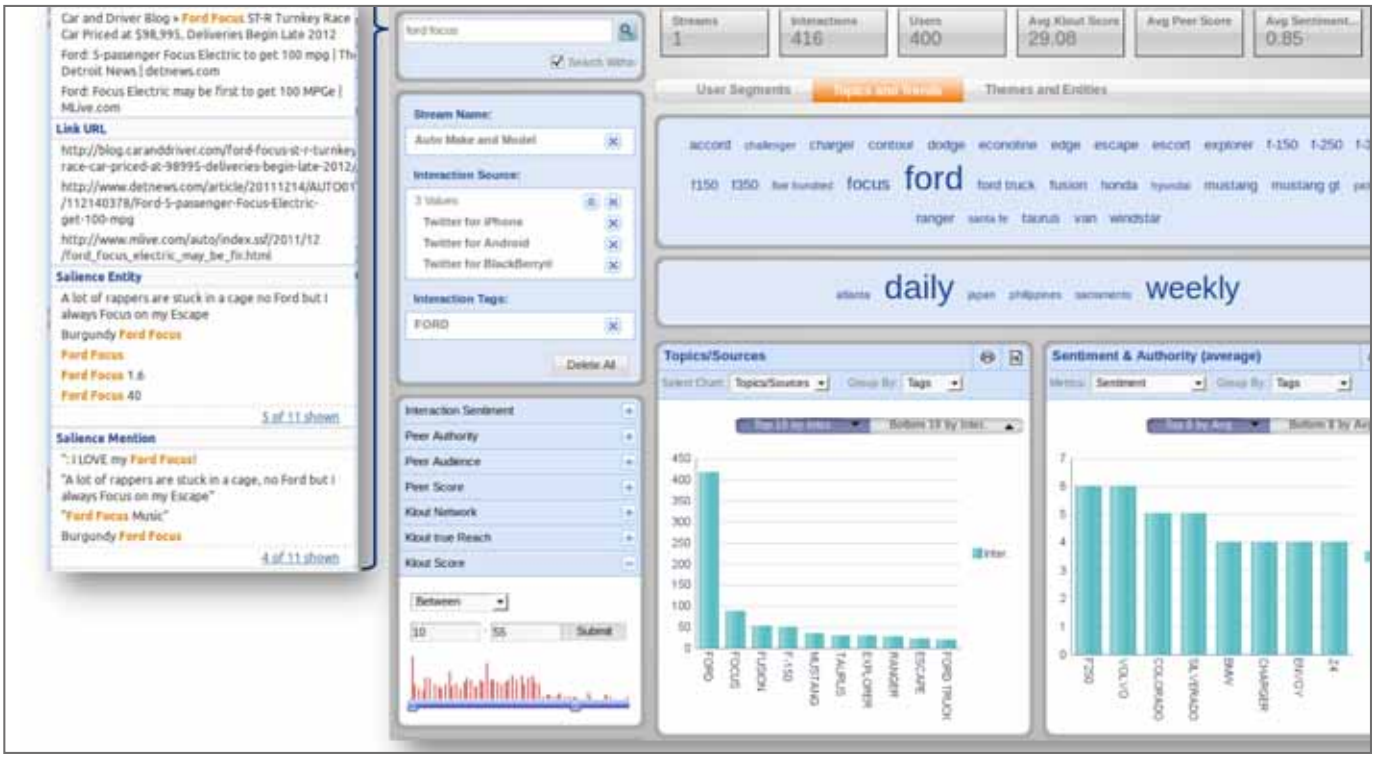


Abbildung 3: Beispiel eines Endeca-Dashboards

Werkzeugen zusätzliche Suchfunktionalitäten, die den unstrukturierten, textuellen Informationen besser gerecht werden. Es handelt sich jeweils um Ergänzungen zum Bestehenden, also eher Evolution als Revolution. Eine technische Kernfrage lautet, wie die unstrukturierten Massendaten aus Big Data mit dem Data Warehouse verbunden werden können. Hierzu gibt es seitens Oracle mehrere technische Möglichkeiten:

- *Oracle Loader for Hadoop*
  Daten aus einem Hadoop-Cluster werden direkt in das Oracle Data Warehouse geladen
- *Oracle Direct Connector for Hadoop HDFS*
  Direkter Zugriff auf das verteilte Filesystem für das Oracle Data Warehouse
- *Oracle Data Integrator (ODI) Application Adapter for Hadoop*
  Einbinden eines Hadoop-Jobs in einen ODI-Ladeprozess

Abbildung 2 zeigt beispielhaft anhand von Twitter-Nachrichten zwei unterschiedliche Szenarien, wie sogenann-

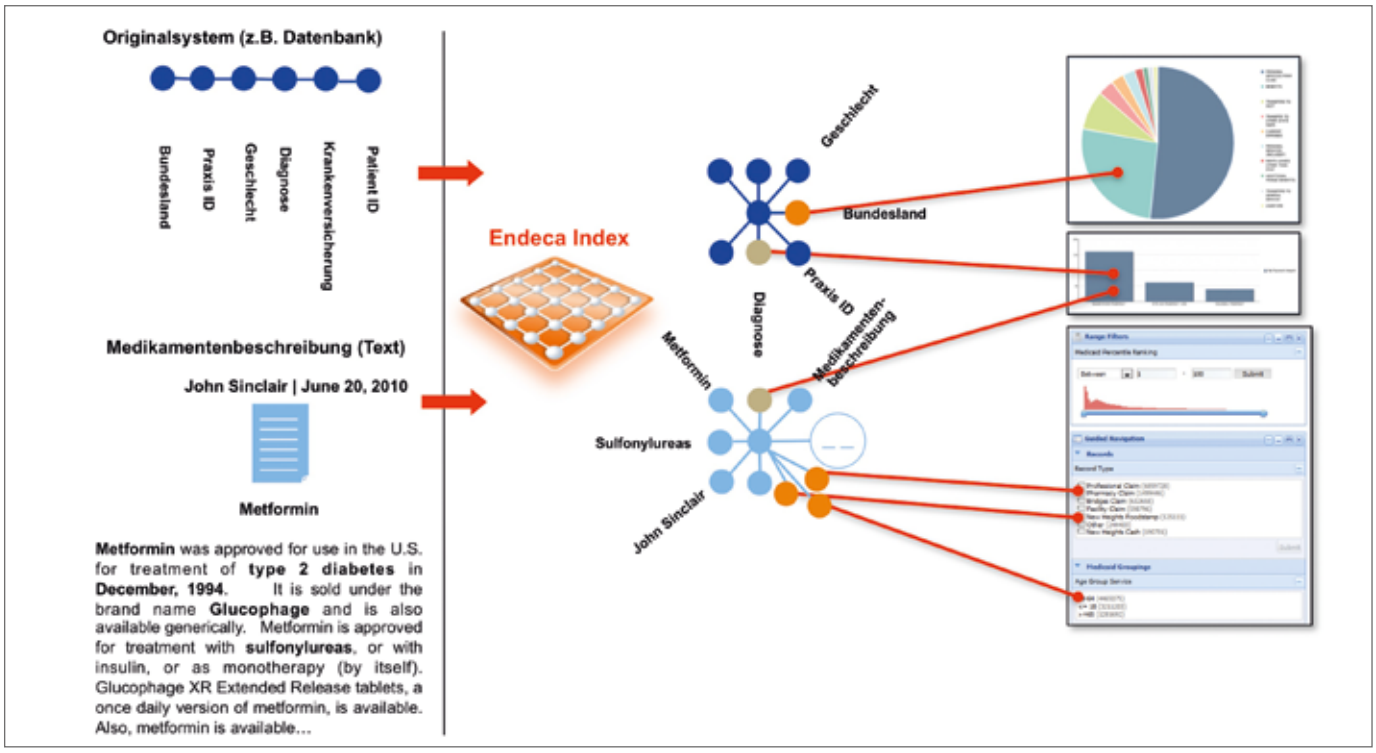


Abbildung 4: Beispiel eines Facetten-Datenmodells

te „Social-Media-Daten" in die Big-Data-Infrastruktur eines Unternehmens überführt und auswertbar gemacht werden können. Szenario 1 steht dabei für den individuellen Entwicklungsansatz, bei dem die Akquisition der Rohdaten über Twitter-Developer-APIs (siehe http://dev.twitter.com) und die Datenorganisation über das Hadoop-MapReduce-Entwicklungs-Framework (nicht abgebildet) erfolgt. Alternativ lassen sich heute auch schon Mehrwertdienste (Szenario 2) in Anspruch nehmen, die per Auftrag Twitter-Datenabzüge aufbereiten und anreichern, indem sie unter anderem den Geo-Bezug herstellen, den Einfluss der Twitter-Beiträge auf andere per „Klout Score" ermitteln oder eine Sentiment-Analyse durchführen. Im Ergebnis werden die relevanten Daten (in der Abbildung die

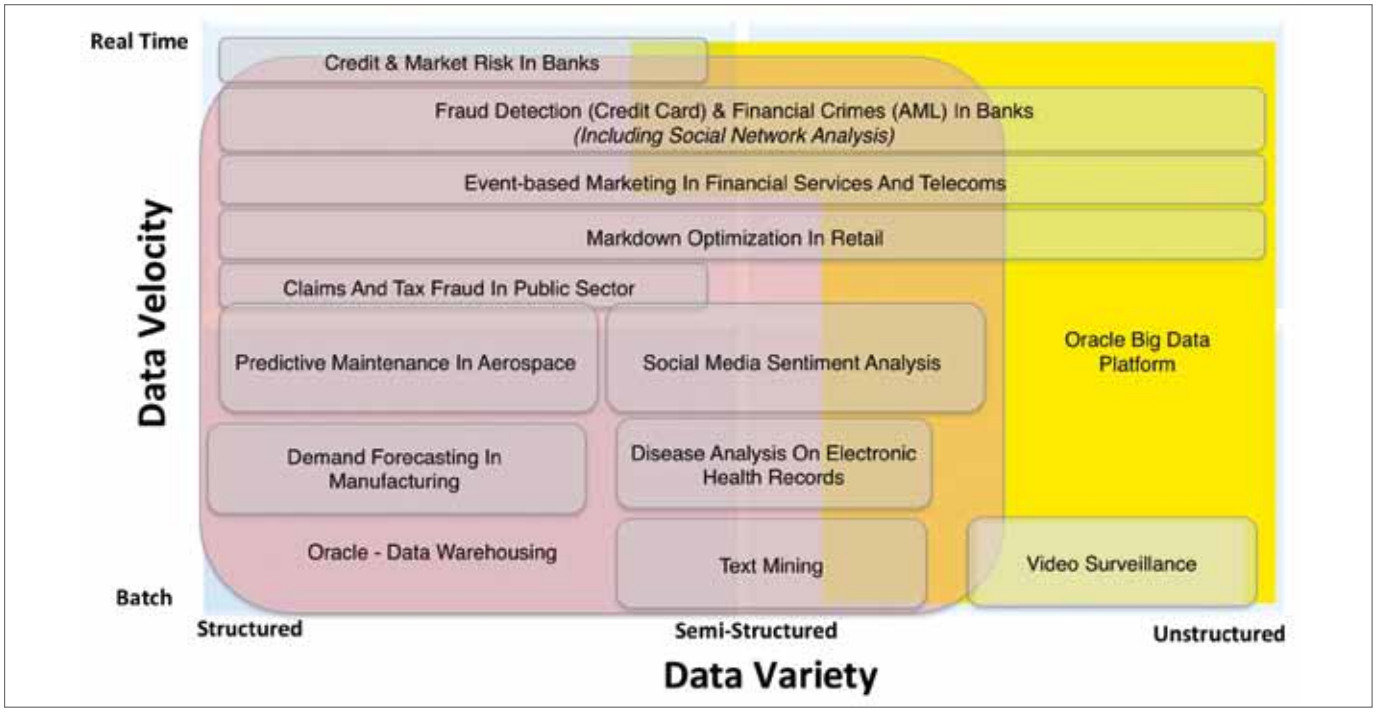


Abbildung 5: Big-Data-Anwendungsbereiche: Oracle Lösungsquadrant

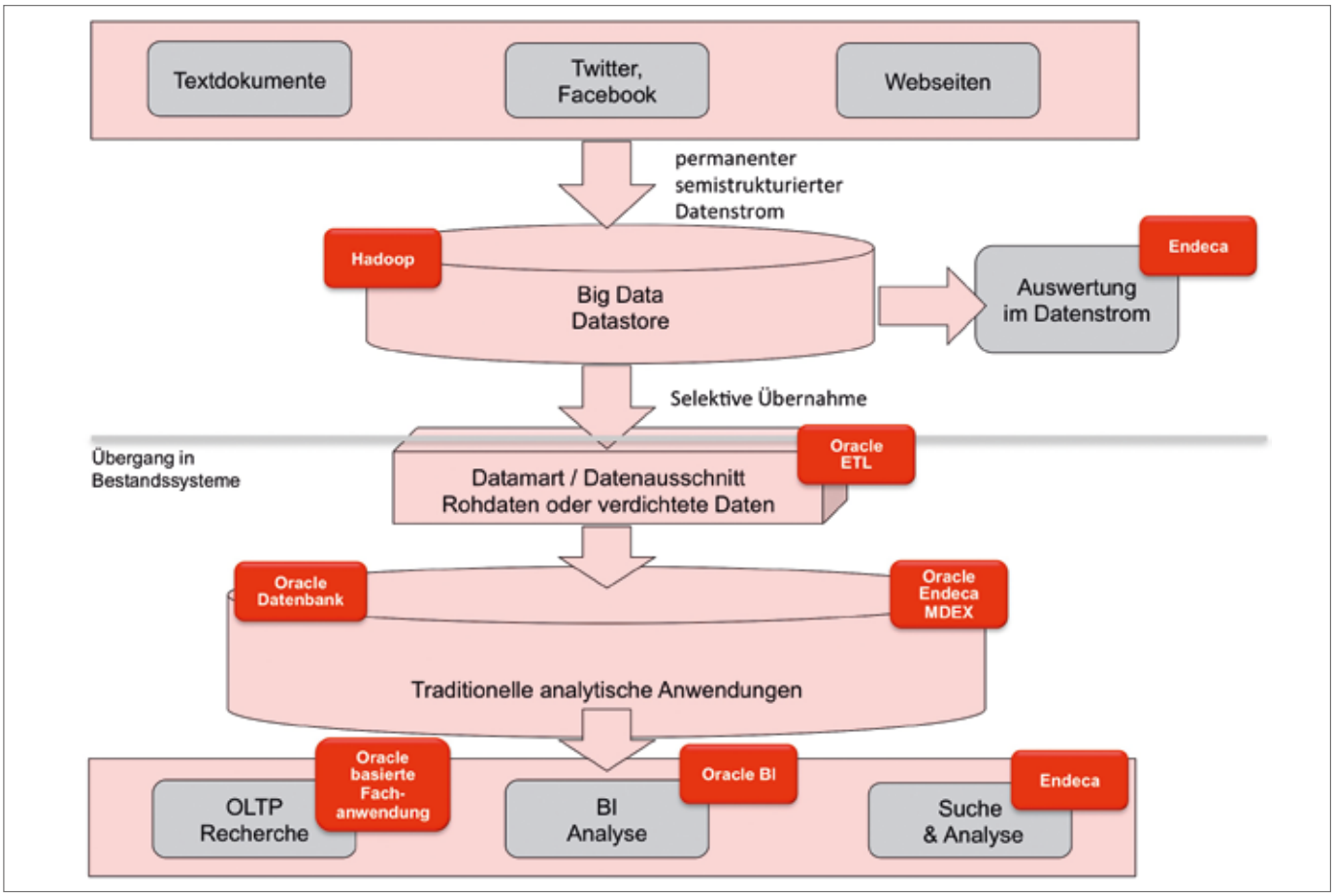


Abbildung 6: Analytisches Gesamtszenario

veredelten Twitter-Feeds) als Key-Value-Paare in Dateiform in einem Hadoop Distributed File System (HDFS) zur weiteren Analyse bereitgestellt. Nutzt man hierzu die Oracle-Big-Data-Infrastruktur in Kombination mit einem Oracle-Data-Warehouse, eröffnet sich dem Analysten ein eleganter Weg des Datenzugriffs per External Tables und SQL (siehe auch [4]).

Analysemöglichkeiten

Die Akquisition von Endeca erweitert das bisherige Oracle-Business-Intelligence-Analyse-Spektrum, indem die textbasierte Suche unstrukturierter Informationen mit den typischen quantitativen BI-Analysen kombiniert und dem Benutzer intuitiv nutzbar präsentiert wird. Die Verbindung quantitativer und qualitativer Informationen überschreitet die klassische Grenze von Business Intelligence und kann konzeptionell dem Knowledge Management zugeordnet werden. Der Slogan „No data left behind“ drückt diese Philosophie treffend aus. Erweiterte Analyse-Funktionen sind zum Beispiel die unternehmensweite Suche, die Präsentation in Form von Tag Clouds, das datengetriebene dynamische Filtern von Merkmalen und die sogenannte „Facetten-Navigation“, bei der die Suche und Auswahl von Attributen wie auf einer Webseite funktioniert [5].

Abbildung 3 zeigt plastisch Teile dieser neuen funktionalen Möglichkeiten. Es geht um die Analyse eines Twitter-Streams zum Thema „Auto Make and Model“. In der Guided-Navigation-Leiste links sieht man die einbezogenen Datenquellen (iPhone-, Android- und Blackberry-Nutzer) und die weiteren gesetzten Filterkriterien („Ford Focus“). Oben in der Metrik-Leiste wird ausgewiesen, dass in 416 (von ca. 350.000 Interaktionen) zutreffende Nachrichten gefunden wurden und sich 400 (der ca. 132.500 Benutzer) zu diesem Thema austauschen. In den Tag-Clouds werden besonders häufig verwendete, unterschiedliche Pkw-Modelle und andere Begriffe hervorgehoben, wobei die Größe der Schrift zeigt, auf welche Wörter die meisten Treffer kommen. Die bereits erwähnten Möglichkeiten zur Anreicherung von Social-Media-Daten durch „Klout Scores“ und Sentiment-Analysen helfen dem Analysten bei der Bewertung der Twitter-Beiträge, etwa in Form zusätzlicher Metriken oder weiterer Attribute für die geführte Suche im Datenbestand. Schließlich finden sich unten weitere Statistiken, die zusätzlichen korrespondierenden Inhalt enthalten können.

Bevor es zur fachlichen Analyse kommen kann, sind die Daten aufzubereiten, gegebenenfalls zu verknüpfen sowie anzureichern. Neben klassischen ETL-Funktionen gibt es seitens Endeca ein erweiterbares Content-Acquisition-System (CAS) für die Daten-Integration von Hunderten von Dateitypen, Dokument-Repositories, CMS-Systemen, Webinhalten und RSS-Feeds. CAS kann sowohl Dateiserver als auch Twitter, Facebook & Co. ana-

lysieren. Jedes unstrukturierte Attribut kann verarbeitet und um weitere Informationen angereichert werden. Gängige Techniken sind:

- Automatic Tagging
- Named Entity Extraction
- Sentiment Analysis
- Term Extraction
- Geospatial Matching

Die unstrukturierten Daten können mit anderen Datensätzen über einen beliebigen Schlüssel miteinander verbunden werden. Natürlich können auch strukturierte Daten mit diesen unstrukturierten Daten im Rahmen des ETL-Prozesses verknüpft sein. Dabei wird keine feste analysefokussierte Datenmodellierung betrieben – wie im Data Warehouse in Richtung Star- oder Snowflake-Modell in Form von fest verknüpften Tabellen üblich –, sondern die Dimensionen werden alle gleichberechtigt nebeneinander in ein Modell gelegt. In der Praxis existieren Analyse-Modelle mit mehreren Hundert Dimensionen. Aus fachlicher Sicht eröffnen sich so unendliche Analyse-Möglichkeiten. Abbildung 4 veranschaulicht die Idee des hochdimensionalen Facetten-Datenmodells.

Die Praxis

Big-Data-Projekte sind kein Selbstzweck. Die neue Technik ist reizvoll, aufgrund des notwendigen Spezialwissens und der sehr großen Datenmengen (Hardware-Bedarf) aber durchaus kostenintensiv. Daher ist es erforderlich, die fachlichen neuen Möglichkeiten, die sich aus Big-Data-Analysen ergeben können, nüchtern zu bewerten. Das kann nur jedes Unternehmen selbst anhand seiner Anwendungsfälle tun. In Anlehnung an [6] zeigt Abbildung 5 eine Gegenüberstellung einiger Big-Data-Anwendungsbereiche und des Oracle-Lösungsangebots zu Big Data und Data Warehousing.

Unter www.doag.org/go/doagnews/erb_tabelle sind beispielhaft fünf ausgewählte Use Cases vorgestellt und ihre Komplexität sowie deren Geschäftsnutzen bewertet.

Quellenverzeichnis

[1] McKinsey Global Institute: Big Data: The next frontier for innovation, competition, and productivity, Report, May 2011: http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

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

[3] Cackett, D./Bond, A./Lancaster,K./Leiker, K., Enabling Pervasive BI through a Practical Data Warehouse Reference Architecture, An Oracle White Paper, Februar 2010: http://www.oracle.com/us/solutions/datawarehousing/058925.pdf

[4] Günther Stürner, Big Data – Hype und Wirklichkeit, Vortrag auf dem Führungskräfte-Forum „Ergebnis- und wirkungsorientierte Steuerung“ des Behördenspiegels: http://www.fuehrungskraefte-forum.de/?page_id=1617

[5] Mark Rittman, Where Does Endeca Fit with Oracle BI and DW?, 22. Februar 2012, http://www.rittmanmead.com/2012/02/oracle-endeca-week-where-does-endeca-fit-with-oracle-bi-dw-and-epm/

[6] Ravi Kalakota, Big Data Analytics Use Cases, 12. Dezember 2011: http://practicalanalytics.wordpress.com/2011/12/12/big-data-analytics-use-cases

[7] TU München, o.V., Neuer Krebsauslöser in Pommes frites entdeckt; scinexx – Das Wissensmagazin, 19. August 2008, http://www.g-o.de/wissen-aktuell-8686-2008-08-19.html

[8] o.V.: Bei Twitter hat Obama im Wahlkampf die Nase vorn, in Westdeutsche Allgemeine Zeitung Online, 3. Januar 2012, http://www.derwesten.de/wirtschaft/digital/bei-twitter-hat-obama-im-wahlkampf-die-nase-vorn-id6210915.html

[9] o.V.: Neue Umsatzsteuer soll Betrug vorbeugen, in Frankfurter Allgemeine Zeitung Online, 20. Oktober 2005: http://www.faz.net/aktuell/wirtschaft/wirtschaftspolitik/haushalt-neue-umsatzsteuer-soll-betrug-vorbeugen-1282102.html

Oliver Röniger
oliver.roeniger@oracle.com

Harald Erb
harald.erb@oracle.com

Wir begrüßen unsere neuen Mitglieder

Persönliche Mitglieder

Norbert Kossok
Dirk Wemhöner
Alexandra Strauß
Thomas Ewald-Nifkiffa
Kevin Brych
Joachim Engel
Thorsten Grebe
Martin Bernemann
Josef Rabacher
Uwe Schreiber
Wolfgang Michael Girsch
Christa Weckman
Thomas Krahn
Marco Stroech
Wolfgang Bossmann
Christoph Mecker
Corinna Kerstan
Gerhard Schaefer
Michael Tucek
Rüdiger Ziegler
Erika Krüger
Andreas Koop
Ulrich Gerkmann-Bartels
Manfred Drozd
Christoph Quereser
Andreas Reinhardt
Markus Vincon

Firmenmitglieder

Dirk Fleischmann, cubus BI Solutions GmbH
Wolfgang Hack, dimensio Informatics GmbH
Volker Oboda, DMySQLAG e.V.
Martin Böddecker, mb Support GmbH
Hans Haselbeck, EMPIRIUS GmbH

In der letzten Ausgabe der DOAG *News* wurden unter der Überschrift „SQL oder NoSQL DB: Das ist die Frage" die Oracle NoSQL Datenbank und (kurz) das quelloffene Hadoop Distributed Filesystem (HDFS) als Systeme zum Speichern großer Mengen un- oder schwach strukturierter Daten mit geringer Informationsdichte vorgestellt. Dieser Artikel baut darauf auf und beschreibt, wie mit den so gespeicherten Daten sinnvoll gearbeitet werden kann.

Einführung für RDBMS-Kenner: Hadoop, MapReduce, Oracle Loader for Hadoop und mehr

Carsten Czarski, ORACLE Deutschland B.V. & Co. KG

Da sowohl NoSQL-Datenbanken als auch das HDFS keine Datenstrukturen kennen und folgerichtig keine Abfragesprache anbieten, ist eine andere Vorgehensweise erforderlich. Abbildung 1 zeigt stark vereinfacht die Einordnung von Big Data in eine IT-Landschaft. Der graue Bereich unten zeigt die traditionelle Vorgehensweise mit strukturierten Daten und hoher Informationsdichte: Daten werden in OLTP-Systemen erfasst, mit ETL-Prozessen in ein Data Warehouse geladen und stehen dann zum Reporting oder für Analyseprozesse zur Verfügung. Für Big Data, also für große Datenmengen mit geringer Informationsdichte, kommen diese Methoden jedoch nicht infrage – allein das Erstellen eines vernünftigen Tabellenmodells für solche Daten wäre schon ein extrem schwieriges Vorhaben (siehe Abbildung 1).

Daher werden diese Daten zunächst in Systemen abgelegt, die ohne Datenmodell auskommen und massiv parallel ausgelegt werden können (siehe Artikel in der letzten Ausgabe). Allerdings bieten diese Systeme folgerichtig keine Abfrage- und Analyse-Möglichkeiten, wie man sie von einem RDBMS her kennt. Zur Auswertung der Daten muss daher prinzipiell der gesamte Datenbestand durchgearbeitet werden – und wegen der riesigen Menge sind hier verteilte Systeme erforderlich.

Verteilte Verarbeitung mit Hadoop

Hadoop ist, kurz gesagt, ein Framework zur verteilten Datenspeicherung und -verarbeitung. Hadoop-Cluster können sehr groß werden. Die größten Installationen haben eine vierstellige Anzahl an Rechnerknoten. Hadoop besteht im Wesentlichen aus zwei Komponenten. Das Hadoop Distributed Filesystem (HDFS) stellt ein über die Rechnerknoten verteiltes Dateisystem bereit. Wie schon im Artikel in der letzten Ausgabe erwähnt, wird dieses zur dateiorientierten Speicherung von Big Data verwendet. Ist dagegen eher die satzorientierte Speicherung gefragt, sind NoSQL-Datenbanken besser geeignet.

Wie ein normales Dateisystem auf einem PC sieht auch HDFS eine Art „File Allocation Table" vor. Diese Datenstruktur enthält die Information, auf welchem Knoten im Cluster die einzelnen Blöcke einer Datei abgelegt sind, und wird auf einem besonderen Rechnerknoten, dem „Name Node", im Hauptspeicher gehalten. Ein Block im HDFS ist allerdings größer als auf einem PC – 256 MB sind der Normalfall. Der „Name Node" weiß, auf welchen „Data Nodes" die Blöcke einer HDFS-Datei liegen. Da in einem verteilten System immer einzelne Knoten ausfallen können, werden die Daten redundant gespeichert. Normalerweise arbeitet man mit einem Replikationsfaktor von drei – jeder Block ist im Cluster also dreimal vorhanden.

Der zweite Teil des Hadoop-Frameworks heißt „MapReduce" und ist ein Programmier-Framework zur ver-

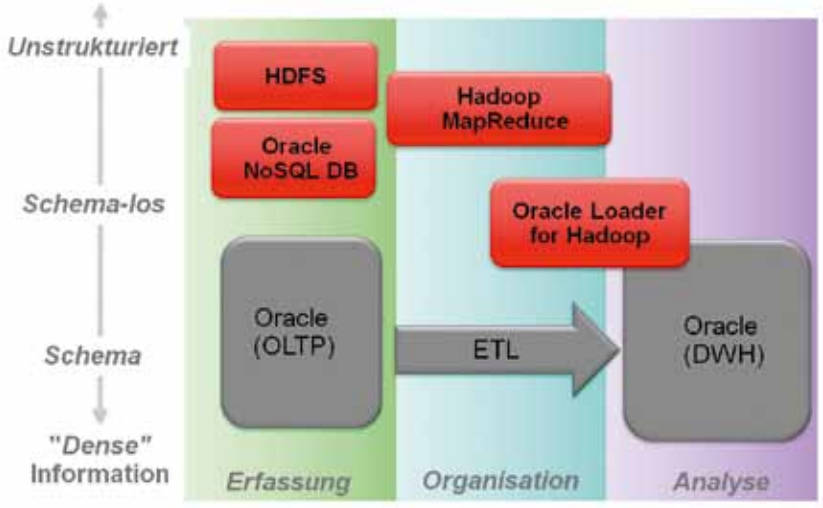


Abbildung 1: Neue Datenformen (Big Data) im Zusammenspiel mit dem Data Warehouse

teilten Verarbeitung. In einem MapReduce-Job sind Code und Datenfluss für die verteilte Verarbeitung optimiert. Der Entwickler eines MapReduce-Jobs kann sich allein auf die Geschäftslogik konzentrieren und muss keinerlei speziellen Code für die verteilte Verarbeitung schreiben.

Abbildung 2 zeigt das Zusammenspiel der Rechnerknoten in einem Hadoop-Cluster. Ein MapReduce-Job wird von einem Client-Rechner an den „Job Tracker" übermittelt. Dieser sorgt dann für die parallele Verarbeitung auf den einzelnen „Data Nodes". Wenn der Job mit Daten im HDFS arbeitet, weist der „Job Tracker" den „Data Nodes" die Teilaufgaben so zu, dass sie – soweit möglich – mit den Datenpaketen arbeiten können, die sie selbst halten. Zwischenergebnisse und temporäre Dateien werden wiederum ins HDFS geschrieben, sodass jeder Rechnerknoten in einem eventuell nachgelagerten Job darauf zugreifen kann.

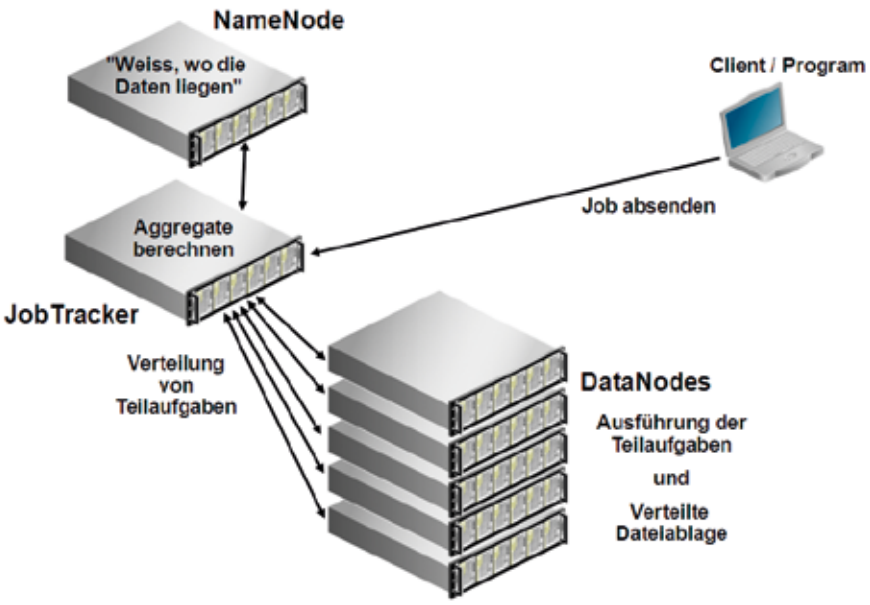


Abbildung 2: Schematische Architektur eines Hadoop-Clusters

MapReduce

MapReduce ist, wie bereits erwähnt, ein Programmier-Framework: Jede Art von Aufgabe kann als MapReduce-Job implementiert und dann verteilt ausgeführt werden. MapReduce schreibt nicht vor, was der Job zu tun hat, sondern vielmehr, wie dieser zu implementieren beziehungsweise wie der Code zu organisieren ist.

Technisch stellt sich das so dar, dass Interfaces vorgegeben sind, die dann vom Entwickler ausprogrammiert werden. Wie immer bei einem Interface sind die Signaturen, also die Ein- und Ausgabeparameter der Funktionen beziehungsweise Java-Methoden, vorgegeben.

In MapReduce-Jobs werden die Daten in Form von Key-Value-Paaren ausgetauscht. Der Entwickler bekommt diese als Eingabe in seinen Job und muss Key-Value-Paare wieder ausgeben. Dabei durchlaufen die Daten stets die folgenden Phasen:

1. Input in den MapReduce-Job
2. Mapper-Phase
3. Shuffle & Sort
4. Reducer-Phase
5. Ausgabe aus dem Job

Zur Veranschaulichung ein Beispiel: Angenommen, in der Oracle NoSQL Datenbank befinden sich 100.000 Key-Value-Paare. Als Schlüssel (Key) dient jeweils die hochgezählte Zahl zwischen 1 und 100.000, als Wert (Value) eine ganzzahlige Zufallszahl zwischen 1 und 100. Die Aufgabe des MapReduce-Jobs ist es nun, für jede Zahl zwischen 1 und 100 deren Vorkommen zu zählen.

```
Key             Value
--------------- -------------
1               67
2               12
3               1
:               :
99999           12
100000          56
```

Listing 1

```
public static class Map
extends Mapper <Text, Text, Text, IntWritable> {
  Text                              key =   new Text();
  private final static IntWritable value = new IntWritable(1);

  @Override
  public void map(Text keyArg, Text valueArg, Context context)
  throws IOException, InterruptedException {
    key.set(valueArg.toString());
    context.write(key, value);
```

Listing 2

```
Input                           Output
Key             Value           Key         Value
--------------- -------------   ----------- -----------
1               67              67          1
2               12              12          1
3               1               1           1
:               :               :           :
99999           12              12          1
100000          56              56          1
```

Listing 3

Eingabe in den MapReduce-Job

Alles beginnt mit der Übergabe der Quelldaten an den MapReduce-Job als Key-Value-Paare. Der Java-Entwickler verwendet hierfür die InputFormat-Klassen. Die vom Hadoop-Framework mitgelieferte Klasse „TextInputFormat" wandelt Dateien im HDFS in Key-Value-Paare um – der Entwickler bekommt dann ein Text-Fragment als „Value" und die Position in der Datei als „Schlüssel". Die Oracle NoSQL Datenbank liefert die Java-Klasse „KVInputFormat" mit, welche die Key-Value-Paare der NoSQL-Datenbank an den MapReduce-Job durchreicht. Für andere Quellsysteme kann man sich eigene InputFormat-Klassen schreiben. Im Beispiel liefert die Oracle NoSQL Datenbank folgende Key-Value-Paare (siehe Listing 1).

```
Key             Value
--------------- -------------
67              {1,1,1,1,...}
12              {1,1,1}
:               :
56              {1,1,1,1...}
```

Listing 4

```
public static class Reduce extends Reducer
  <Text, IntWritable, Text, IntWritable> {
  private IntWritable result = new IntWritable();

  public void reduce(
    Text key, Iterable<IntWritable> values, Context context
  ) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {sum += val.get();}
    result.set(sum);
    context.write(key, result);
  }
```

Listing 5

```
Input                          Output
Key             Value          Key         Value
--------------- -------------  ----------- -----------
67              {1,1,1,1,...}  67          617
12              {1,1,1}        12          3
1               {1,1,1,1,...}  1           132
:               :              :           :
56              {1,1,1,1...}   56          872
```

Listing 6

Mapper

Der Mapper ist eine eigene Java-Methode (genauer: eine eigene Java-Klasse) mit definierter Ein- und Ausgabeschnittstelle. Hier führt der Entwickler ein „Mapping" durch, indem er die empfangenen Key-Value-Paare auf neue Key-Value-Paare abbildet (siehe Listing 2).

Wichtig ist, dass der Entwickler zur gleichen Zeit stets nur ein Key-Value-Paar sieht. Somit kann ein Zusammenfassen zu diesem Zeitpunkt noch nicht erfolgen, wohl aber kann (und muss) der Entwickler dieses hier vorbereiten, indem er die Key-Value-Paare, die zusammengefasst werden sollen, auf neue Key-Value-Paare mit gleichem Schlüssel abbildet. Im Beispiel würde der Mapper die empfangenen Key-Value-Paare also wie in Listing 3 zeigen.

Da gezählt werden soll, wie oft jede Zufallszahl vorkommt, müssen die Key-Value-Paare anhand der Zufallszahl zusammengefasst werden – der Mapper legt also die Zufallszahl als neuen Schlüssel fest. Als Wert dient die Zahl „1" – denn die Zufallszahl ist „einmal vorgekommen". Wie schon gesagt, sieht der Mapper nur ein Key-Value-Paar auf einmal. Das ist die Grundlage für die verteilte Verarbeitung. Die eigentliche Zählung erfolgt später im Reducer.

Shuffle & Sort

Diese Funktion wird vom Hadoop-Framework übernommen – der Entwickler muss nichts machen. Es werden die vom Mapper ausgegebenen Key-Value-Paare nach Schlüsseln sortiert, zusammengefasst und dann an den ebenfalls vom Entwickler implementierten Reducer übergeben. Im Beispiel sieht das aus wie in Listing 4.

Shuffle & Sort fasst die 100.000 Key-Value-Paare zusammen, die der Mapper ausgegeben hat, sortiert sie und ordnet sie Schlüsseln zu. Ein Schlüssel (die Zufallszahl) kommt nun nur noch einmal vor und enthält alle Werte (im Beispiel nur 1er) als Array. In dieser Form gehen die Key-Value-Paare an den Reducer.

Reducer

Wie der Mapper ist auch der Reducer eine eigene Java-Methode mit einer definierten Eingabe- und Ausgabe-Schnittstelle. Der Entwickler bekommt die Schlüssel und Werte-Arrays aus der „Shuffle & Sort"-Phase übergeben und kann diese nun weiterverarbeiten. Dabei erzeugt er wiederum Key-Value-Paare aus der Ausgabe. „Zählen" bedeutet dabei das Durcharbeiten des Arrays und gleichzeitige Hochzählen einer Variable wie bei „cnt = cnt + 1". Für das Beispiel sieht der Code des Reducer wie in Listing 5 aus. Dieser Code wandelt die Key-Value-Paare um (siehe Listing 6).

Ausgabe aus dem MapReduce-Job

Der letzte Schritt ist wiederum die Ausgabe der vom Reducer generierten Key-Value-Paare. Verwendet der Entwickler die vom Hadoop-Framework mitgelieferte Klasse „TextOutputFormat", so werden die Key-Value-Paare als Textdatei im HDFS abgelegt. Die ebenfalls mitgelieferte Klasse „SequenceFileOutputFormat" legt die Paare im Binärformat im HDFS ab. Das ist sinnvoll, wenn ein nachgelagerter MapReduce-Job damit weiterarbeiten soll. Und na-

türlich kann ein Entwickler auch eigene OutputFormat-Klassen schreiben. Für gängige Aufgaben haben sich in der OpenSource Community mittlerweile Standard-Implementierungen durchgesetzt – man muss also nicht wirklich alles selbst machen. Einige Implementierungen sind nachfolgend kurz vorgestellt.

Hive

Hive ist ein sehr mächtiger, generischer MapReduce-Job, der es erlaubt, SQL-Abfragen auf Dateien im HDFS auszuführen. Auf den ersten Blick fragt man sich, wie das sein kann, denn es wurde ja schon mehrfach gesagt, dass NoSQL-Datenbanken und das HDFS eben keine SQL-Abfragesprache anbieten und das wegen fehlender Strukturen auch gar nicht können. Hive erlaubt die Definition einer Tabelle auf Basis einer Datei im HDFS – diese Tabelle wird ähnlich zu einer externen Tabelle in der Oracle-Datenbank definiert und funktioniert auch ganz ähnlich. Ist die Tabelle definiert, so kann sie mit SQL abgefragt werden – und Hive arbeitet hier wiederum ganz ähnlich wie die Oracle-Datenbank mit externen Tabellen. Die SQL-Abfrage wird geparst, es wird eine Art „Ausführungsplan" erstellt und die Ausführung erfolgt als MapReduce-Job, der die HDFS-Datei Zeile für Zeile durcharbeitet und dabei die in der SQL-Abfrage ausgedrückte Logik abarbeitet. Die Abbildung der Textdatei auf Zeilen und Spalten sowie das Filtern anhand der WHERE-Klausel erfolgt in der Mapper-Phase, Aggregats-Funktionen werden im Reducer abgebildet.

Enthält das HDFS Dateien von wohlbekannter Struktur (Log-Dateien), auf der man eine externe Tabelle definieren kann, so erlaubt Hive die Auswertung derselben mit SQL – was wesentlich einfacher und schneller ist als das Selbst-Schreiben eines MapReduce-Jobs. Allerdings wird nach wie vor der ganze Datenbestand durchgearbeitet. Auf Daten in der NoSQL-Datenbank kann Hive (noch) nicht arbeiten.

Sqoop

Diese Funktion bietet das Laden einer Datei im HDFS in eine relationale Datenbank per JDBC. Wie für Hive muss die HDFS-Datei auch für Sqoop eine gewisse Struktur aufweisen und es ist gegebenenfalls ein Mapping auf Spalten-Namen in einer Tabelle nötig. Sqoop parst die Datei, generiert die Spalten und Zeilen in der Mapper-Phase und schreibt sie in der Reducer-Phase per JDBC in die Datenbank. Analog dazu ist auch der umgekehrte Weg, also das Auslesen einer Datenbank-Tabelle per JDBC und die Ablage der Daten als Textdatei im HDFS, möglich.

Oracle Loader for Hadoop

Als Alternative zu Sqoop bietet Oracle als Teil der „Big Data Connectors" den „Oracle Loader for Hadoop" an. Im Gegensatz zum generischen Sqoop ist dieser für die Oracle-Datenbank optimiert und eignet sich besonders für das Laden großer Datenmengen in diese. Oracle Loader for Hadoop ist ebenfalls als MapReduce-Job implementiert und erlaubt zusätzlich zum Laden per JDBC auch das Batch-orientierte Laden in die Datenbank. Dafür kann beispielsweise eine Data-Pump-Datei erzeugt werden, die dann als externe Tabelle (Data-Pump-Format) in die Oracle-Datenbank eingebunden werden kann. Die Übernahme in die Zieltabellen kann dann mit den Mitteln der Datenbank (INSERT ... SELECT, Multi-Table INSERT oder Pipelined Functions) erfolgen. Gerade bei großen Datenmengen ist dieses Verfahren wesentlich effizienter.

Fazit

Letztlich hat sich der Kreis geschlossen: Big Data sieht zunächst vor, dass große Mengen un- oder nur schwach strukturierter Daten als Dateien ins HDFS oder als Key-Value-Paare in die Oracle NoSQL Datenbank gespeichert werden. Diese verteilten Systeme sind in der Lage, auch größte Datenmengen mit kurzen Antwortzeiten aufzunehmen und ständig zu wachsen.

Da hier keine Abfragesprache existiert, muss für jede Auswertung der gesamte Datenbestand durchgearbeitet werden. Diese Aufgabe übernimmt MapReduce – ebenfalls massiv parallel. Die gewünschten Auswertungen beziehungsweise Aggregationen werden als MapReduce-Jobs implementiert und im Hadoop-Cluster ausgeführt. Als letzter Schritt einer solchen Jobkette kann schließlich der Oracle Loader for Hadoop ins Spiel kommen, der die gefundenen Aggregate in die Oracle-Datenbank lädt, wo sie Teil des Data Warehouse und damit zur Basis für Reporting, Business Intelligence und weitere Analyse werden können.

Big Data Appliance

An dieser Stelle bietet sich die Einordnung der „Oracle Big Data Appliance" an. Dieses Engineered-System ist speziell für die Anforderungen eines Hadoop-Clusters oder der NoSQL-Datenbank ausgelegt. Im Gegensatz zur sehr aufwändigen Einrichtung eines Hadoop-Clusters mit Standard-Hardware findet der Setup einer Big Data Appliance skriptgesteuert in kürzester Zeit statt. Damit ist dieses Engineered-System hochinteressant, wenn man Big-Data-Technologien nutzen möchte, die Installation und den Betrieb eines Hadoop- oder NoSQL-Datenbank-Clusters auf „Commodity-Hardware" jedoch vermeiden möchte.

Weitere Informationen

[1] Oracle Dojo, Eine Einführung in Big Data: http://www.oracle.com/webfolder/technetwork/de/community/dojo/index.html
[2] Whitepaper Big data Overview: http://www.oracle.com/technetwork/server-storage/engineered-systems/bigdata-appliance/overview/wp-bigdatawithoracle-1453236.pdf
[3] Apache Hadoop (Map Reduce und HDFS):http://hadoop.apache.org
[4] Oracle NoSQL Datenbank im OTN: http://www.oracle.com/technetwork/products/nosqldb/overview/index.html

Carsten Czarski
carsten.czarski@oracle.com
http://twitter.com/cczarski
http://sql-plsql-de.blogspot.com

„Alles nur geklaut" titelte die deutsche Band „Die Prinzen" Mitte der neunziger Jahre ihre CD. Wer hätte damals gedacht, dass dieser Titel zwanzig Jahre später in leicht veränderter Schreibweise zu neuer Aktualität gelangen würde?

„Alles nur gecloud …"

Sven Kinze und Martin Verleger, Apps Associates GmbH

Wer heute die einschlägigen Gazetten durchblättert, kommt an dem Thema „Cloud Computing" nicht vorbei. Bezogen auf die Vielzahl der Veröffentlichungen und Events entsteht langsam der Eindruck, dass es im Bereich des professionellen Datenbetriebs nicht mehr ohne Cloud-Anwendungen geht. Die Wirklichkeit sieht anders aus: Im Gegensatz zu anderen Ländern wie den USA fristet diese Form des Systembetriebs in Deutschland noch ein Schattendasein. Als mittelständisches Unternehmen, das auf beiden Seiten des Atlantiks aktiv ist, treibt auch Apps Associates die Frage nach den deutschen Vorbehalten gegen diesen Ansatz um. Eine einhellige Antwort hierfür haben die Autoren nicht – was sie aber nicht daran hindert, sich mit dem Thema zu befassen.

Die Theorie

Im Grundsatz geht es bei Cloud Computing darum, seine IT-Infrastrukturen einem zunehmend dynamischen Bedarf anzupassen. Betriebswirtschaftlich wird hier die alte „Make-or-buy-Frage" erneut thematisiert. Im Cloud-Zeitalter ist Fremdbeschaffung Trumpf. Interessant ist dieses Vorgehen überall da, wo es wirtschaftlich sinnvoll ist, Dinge auszulagern, und wo keine zwingenden Gründe wie etwa aus dem Umfeld „Governance, Risk und Compliance" dagegenstehen.

In Zeiten des Fachkräftemangels kann jede Betriebs-IT sehr schnell in die Situation kommen, Infrastrukturen oder deren Teile nicht mehr selbst betreiben zu können. Dies gilt vor allem in besonderen Belastungssituationen, wenn zum Beispiel neue Anwendungen getestet werden sollen, ohne dass die hierfür nötige physikalische Infrastruktur zur Verfügung steht.

Überhaupt scheint die Cloud im Falle von Schnellschüssen ein Tausendsassa zu sein: Man denke an globale Anwenderschulungen, an die Sicherung von Systemzuständen anlässlich bestimmter Meilensteine, an das schnelle Klonen von ganzen Systemen, um für eventuelle Fehlersuche ein Referenzsystem zu haben, oder schlicht und ergreifend an die gute alte „Sandbox", in der man beim Experimentieren auch mal ungestraft Fehler machen darf.

Die Praxis

Wie jedoch nähert man sich dem Thema in der Praxis, ohne finanzielle Risiken einzugehen? Zunächst ist es wichtig, ein konkretes Projekt vor Augen zu haben. Ein von Apps Associates bereits häufig durchdekliniertes Szenario ist das Aufsetzen einer Oracle Business Intelligence Suite in der Cloud. Hier kann es zwei mögliche Ausprägungen geben: Entweder das klassische Oracle-BI-Entwicklungsprojekt, bei dem entlang des gesamten Technologie-Stacks alle notwendigen Komponenten eigens ausprogrammiert werden, oder das vorgefertigte Produkt Oracle BI Analytics.

Bei der zweiten Lösung, die sich insbesondere in Nordamerika großer Beliebtheit erfreut, werden sämtliche BI-Komponenten (ETL, Data Warehouse, Metadaten und Reports) als „Out-of-the-Box"-Lösung geliefert – und dies für viele gängige Business-Applikationen wie SAP, Oracle E-Business Suite, Oracle Siebel CRM und andere. In beiden Fällen handelt es sich um einen komplexen Technologie-Stack. Wer nun plant, diese Anwendung als Cloud-Applikation zu betreiben oder zu testen, sollte sich der Hilfe eines erfahrenen Partners bedienen, denn es gibt zwei Herausforderungen: die Cloud einerseits und die anspruchsvolle Welt des OBI andererseits.

Zweitens muss die Frage nach dem passenden Cloud-Provider beantwortet werden. Hier gibt es mittlerweile eine große Auswahl, vor der auch Apps Associates stand. Es wird wahrscheinlich viele Kunden des Online-Versandhauses amazon.com überraschen zu erfahren, dass das Versandhaus gleichzeitig einer der größten Cloud-Anbieter weltweit ist. Im Rahmen einer Gesamtstrategie „Amazon Web Services (AWS)" werden zum Beispiel Services unter dem Namen „Amazon Elastic Computing Cloud (EC2)" angeboten. Hier sind drei Nutzungs- und Preismodelle denkbar: Beim „On Demand"-Ansatz werden nur die tatsächlich konsumierten Ressourcen berechnet. Beim Modell „Reserved" zahlt der User eine Einmalgebühr und dafür erheblich geringere verbrauchsbezogene Nutzungsentgelte. Beim Modell der „Spot Instances" richtet sich der Preis nach Angebot und Nachfrage. In Zeiten geringer Auslastung der gesamten Amazon-Cloud ist es möglich, günstig System-Ressourcen zur zeitweisen Nutzung zu erwerben. Die Preise hierfür ergeben sich in einer Art Versteigerungsverfahren.

Neben der Entscheidung für ein Preismodell müssen – wie in jedem anderen Infrastrukturprojekt auch – viele andere Dinge berücksichtigt werden. Eine entscheidende Frage ist beispielsweise die nach dem sogenannten „Instanztyp". Hier bietet EC2 sechs Grundtypen an, die man dem Sizing und den Erfordernissen seiner Anwendung anpassen muss. Mag der Instanzentyp „Standard" unter Umständen für eine kleine CRM-Anwendung reichen, so kann eine speicherintensivere

Anwendung wie ein Data Warehouse gerne auch schon einmal eine „High-Memory-Instanz“ notwendig machen. Die Wahl des Betriebssystems obliegt dem Kunden.

Die System-Konfiguration erfolgt über das Web mit einem einfachen Firefox-Browser, der ein simples Administrations-Plug-in benötigt. Ein Setup-Wizard leitet den Administrator durch den standardisierten Prozess. Nach der Wahl des Instanzentyps kann die Instanz auf die Bedürfnisse des Projekts angepasst werden. Es folgen Einstellungen zur Systemsicherheit sowie zum Instanzen-Namen. Der Administrator legt die initialen Laufwerkskapazitäten fest und ordnet die Laufwerke der Instanz zu. Den Abschluss bildet die Konfiguration der Firewall. Damit ist der Rechner konfiguriert und kann für das Projekt genutzt werden. Parallel zum gesamten Setup ermittelt ein Konfigurations-Tool die Kosten der virtuellen Maschinen.

In unserem OBI-EE-Beispiel hat sich der Anwender für eine zweistufige Architektur mit zwei großen Maschinen auf Oracle-Linux-Basis entschieden, die jeweils über zwei CPUs, 7.5 GB RAM und 500 GB Plattenplatz verfügen. Die Datenhaltung erfolgt dabei in sogenannten „Elastic Block Stores (EBS)“-Volumes unabhängig von der eigentlichen Instanz.

Die Installation des OBI EE erfolgt wie gewohnt: Auf der ersten Maschine wird eine Oracle-11*g*-R2-Datenbank aufgesetzt, die zweite Maschine beherbergt den WebLogic-Server mit den OBI-11*g*-Komponenten. Zunächst bekommen die Maschinen jeweils eine sogenannte „EC2 Private IP Address“, sodass sie innerhalb der Cloud erreichbar sind und untereinander kommunizieren können. Ziel ist es jedoch, die Server von überall zu erreichen. Eine kleine Besonderheit in der Cloud besteht darin, dass für diesen Schritt beide Services gestoppt werden müssen. Aus Erfahrung ist es ratsam, den nunmehr frischen Stand seiner Cloud-Anwendungen zu sichern. Mit Amazon Machine Image (AMI) bietet EC2 ein einfaches Handling für die Erstellung von Images, die auch für den Download auf die heimischen Festplatten zur Verfügung stehen und somit das flüchtige Dasein ihrer virtuellen Hardware leicht überleben können. Mit der Zuweisung einer sogenannten „Elastic IP“ ist die Installation beendet. Dieses Verfahren, vergleichbar mit NAT, stellt sicher, dass die Cloud-Instanz von außerhalb der Cloud erreichbar ist, und kann einfach auch auf eine andere Instanz übertragen werden. Nun stehen beide virtuelle Maschinen für das Projekt, den Testlauf, das Prototyping oder für die Demo zur Verfügung.

Fazit

Cloud Computing ist keine Raketenwissenschaft. Wer sich damit beschäftigen möchte, ohne heute schon echten Handlungsdruck zu verspüren, der möge sich ein passendes Projekt, einen Vor-Ort-Partner und einen Cloud-Anbieter suchen, um erste Erfahrungen zu sammeln und einfach loszulegen. Denn eines ist sicher – in naher Zukunft wird nicht alles, aber vieles „gecloud“ sein.

Sven Kinze
sven.kinze@appsassociates.com

Martin Verleger
martin.verleger@appsassociates.com

Newsticker

Neu: Identity Management 11*g* Release 2

Die neue Version bietet einfachen und sicheren Zugang für Social-Web-Anwendungen, die Cloud und mobile Umgebungen. Sie ist ein wesentlicher Baustein von Oracle Fusion Middleware. Insgesamt umfasst Oracle Identity Management 11*g* Release 2 das komplette Identity-Management-Portfolio, das in die drei Bereiche Identity Governance, Access Manager und Directory Services aufgeteilt ist.

Der neue Privileged Account Manager ist eine Art „Self-Service-Einkaufswagen“, um Zugang zu neuen Anwendungen zu beantragen. Zu den neuen Funktionen gehören auch native Sicherheit und natives Single-sign-on für mobile Endgeräte sowie Unterstützung für Social-Single-sign-on via Facebook, Google, Yahoo, Twitter und LinkedIn. Directory Services erlauben häufige Updates in dem Verzeichnis, wie sie von standortabhängigen Diensten (Location Based Services) gefordert werden, um mobile und soziale Anwendungen zu unterstützen. Die neue Version enthält außerdem eine „Optimized Solution“ für Oracle Unified Directory mit erhöhter Skalierbarkeit und Zuverlässigkeit, wie sie für Cloud-, mobile und soziale Umgebungen erforderlich ist. Zudem wurde die Flexibilität für den Einsatz in großen Unternehmen und Anwendungsumgebungen erhöht. Erreicht wird dies durch eine Vereinheitlichung von Storage, Proxy, Synchronisierung und Virtualisierung. Dadurch wird die Verwaltung und Installation vereinfacht sowie die Interoperabilität mit einer Vielzahl von Hardware und Betriebssystemen sichergestellt.

Das Partitionieren von Tabellen ist ein Allroundmittel im Data-Warehouse-Umfeld. Die Vorteile reichen von der Administration bis zur Abfrage-Performance, und auch beim Laden ist der Nutzen groß. Der Artikel gibt praktische Hinweise beim Einsatz von Partitionierung.

Sieben gute Gründe für den Einsatz von Partitionierung im Data Warehouse

Detlef E. Schröder und Alfred Schlaucher, ORACLE Deutschland B.V. & Co. KG

In Data-Warehouse-Systemen können einzelne Tabellen in den Bereich von vielen Milliarden Sätzen anwachsen. Dies birgt verschiedene Herausforderungen, die trotz rasanter Weiterentwicklung der Hardware auch softwaretechnisch behandelt werden sollten.

Das Partitionieren großer Tabellen gehört mit zu den wichtigsten Hilfsmitteln des Oracle Data Warehouse. Es teilt die Daten großer Tabellen in physikalisch separierte Mengen und spricht diese Teilmengen separat an. Dadurch müssen bei entsprechenden Abfragen anstelle der gesamten Tabelle nur die Partitionsdaten gelesen werden, was zu einer besseren Abfrage-Performance führt. Darüber hinaus sind die kleineren Datenpakete eine handlichere Verarbeitungseinheit für die Verwaltung des Data Warehouse. Die Gesamtsicht auf alle Daten bleibt dennoch erhalten, sodass die bestehenden Abfragen oder Applikationen nicht verändert werden müssen. Zusätzlich entstehen durch die Aufteilung einzelne, definierte und physisch aufgeteilte Datensegmente (siehe Abbildung 1).

Abbildung 1: Partitionierte Tabellen haben mehrere Sichten. Die Tabelle kann als Ganzes oder in einzelnen Partitionen angesprochen werden.

Der Nutzen des Partitioning-Features einer Datenbank wird oft nur auf die Performance-Verbesserung reduziert. Dies trifft die Möglichkeiten und Vorteile aber nur zu einem, wenn auch wichtigen Teil. Insgesamt gesehen gibt es folgende sieben Gründe, die den Einsatz von Partitionierung empfehlen:

- Abfrage der Performance
- Unterstützung im Ladeprozess
- Vereinfachte Handhabung bei der Indizierung
- Vereinfachtes Update der Materialized View
- Unterstützung bei der Komprimierung
- Unterstützung im Backup-Prozess
- Unterstützung des Information Lifecycle Managements

Abfrage der Performance

Oft wird nur ein Teil der Daten für bestimmte Abfragen benötigt (die letzte Ladeperiode, der letzte Monat, das letzte Quartal etc.). Wenn die Daten in für die Abfrage passenden Häppchen vorliegen, also für jede Ladeperiode eines, dann müssen für eine Abfrage nach der aktuellen und der Vorperiode nur zwei Teilmengen gelesen werden und nicht wie im unpartitionierten Zustand alle 36 vorgehaltenen Ladeperioden. Dadurch wird die zu lesende und zu verarbeitende Datenmenge eingeschränkt und es entsteht ein Performance-Vorteil für die Abfrage. Darüber hinaus können Tabellen, die nach den gleichen Kriterien aufgeteilt sind, auch nur über die notwendigen Teilmengen gejoint werden, was ebenfalls einen Verarbeitungsvorteil erzeugt und die Performance steigert.

Wie werden nun die Inhalte auf die Partitionen aufgeteilt und welche Varianten gibt es? Ein wichtiges Merkmal der Partitionierung ist der Partition Key. Das ist das Feld in der Tabellen-Struktur, über dessen Inhalt das System die Zuordnung von Sätzen zu einer bestimmten Partition festlegt. Damit ist ein sehr praktisches Feature des Partitioning benannt: das System, das bei einem INSERT einen Satz automatisch in die passende Partition aufnimmt. Derjenige, der den ETL-Prozess steuert, muss sich also darum nicht kümmern.

Die Partition-Variante (Range, List, Hash, System, Virtual Column) bestimmt den Umgang mit dem Partition Key. Die am häufigsten genutzte Partitioning-Variante ist das Rang-Partitioning, bei dem die Einteilung der Partitionen entlang definierter Wertebereiche erfolgt. Das im Data Warehouse am häufigsten genutzte Kriterium ist die Zeit (Tage, Monate etc.), also ein Feld mit einem Datumswert (siehe Listing 1).

Andere Partitionierungs-Kriterien können auch zusammengesetzte Felder, über Funktionen ermittelte Werte oder nach dem Alphabet sortierte Werte sein. Diese Varianten lassen sich über Sub-Partitioning in einer zweiten Ebene mischen, um das Partitionieren noch flexibler zu gestalten. Das grundlegende Partitionierungskriterium ist beispielsweise eine Monatseinteilung (Range); die Monatspartitionen können noch weiter nach Verkaufsgebieten (List) unterteilt sein. Mit Sub-Partitioning gewinnt man ein zusätzliches

```
CREATE TABLE F_bestellung_part_range (
SUMME NUMBER(14,0),
MENGE NUMBER(14,0),
BESTELLDATUM DATE,
FK_ARTIKEL_ID NUMBER,
FK_KUNDEN_ID NUMBER,
FK_ORT_ID NUMBER,
FK_DATUM_ID NUMBER,
auftragsart VARCHAR2(30))
PARTITION BY RANGE (bestelldatum) (
PARTITION jan11 VALUES LESS THAN (TO_DATE(,2011-02-01‘, ,SYYYY-MM-
DD‘)) TABLESPACE DWH_SPINDEL,
…
PARTITION feb12 VALUES LESS THAN (TO_DATE(,2012-03-01‘, ,SYYYY-MM-
DD‘)) TABLESPACE DWH_SPINDEL,
PARTITION next_month VALUES LESS THAN (MAXVALUE) TABLESPACE DWH_SPIN-
DEL);
```

Listing 1

```
CREATE TABLE F_BESTELLUNG
SUMME NUMBER(14)
MENGE NUMBER(14)
BESTELLDATUM DATE
FK_ARTIKEL_ID NUMBER
FK_KUNDEN_ID NUMBER
FK_ORT_ID NUMBER
FK_DATUM_ID NUMBER
AUFTRAGSART VARCHAR2(30) ... ;
```

Listing 2

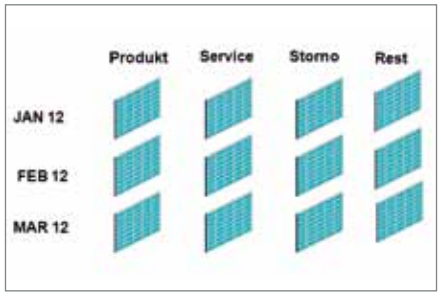


Abbildung 2: Sub-Partitioning „Range/List“

```
select sum(summe) from F_BESTELLUNG_[xxx] where Auftragsart = ‘SERVICE’
and BESTELLDATUM = to_date(‘10.10.2011’,’DD.MM.YY’);
```

Listing 3

Tabelle	Blocks	GB	Millionen Zeilen	Abfragezeit
F_Bestellugen_X	2431440	19,45	411	314,3 sec
F_Bestellugen_Range	2429231	19,43	411	17,3 sec
F_Bestellugen_Range _List	2463418	19,71	411	5,6 sec

Tabelle 1

fachliches Abfragekriterium zu dem der Datengewinnung oder -entstehung – hier der Zeit.

Ein einfacher Test zeigt bereits die Performance-Effekte des Partitioning und Sub-Partitioning. Die Beispiel-Faktentabelle wird in drei Varianten unterteilt: nicht partitioniert, partitioniert (Range) und sub-partitioniert (Range/List) (siehe Listing 2). Listing 3 und Tabelle 1 zeigen die Beispielabfrage auf die Daten und die entsprechenden Laufzeiten.

Wie deutlich zu erkennen ist, liefert die sub-partitionierte Variante das beste Ergebnis, noch vor der einfachen Partitionierung. Wie bereits gesagt: Im Data Warehouse ist Partitionierung bei großen Tabellen ein Muss.

Partition Wise Join

Der Zugriff auf ein Data Warehouse erfolgt meist durch ein Starschema, das aus mehreren Dimensionen und Fakten besteht. Dies bewirkt bei einer Abfrage eine Reihe von Joins zwischen einer meist sehr großen Fakten-Tabelle und mehreren Dimensions-Tabellen. Ist die Fakten-Tabelle partitioniert, kann eine Abfrage mit einem Join zwischen Fakten mehrerer Dimensions-Tabellen in mehrere kleine Joins zergliedert werden. Für jede Partition der Fakten-Tabelle wird ein eigener Join mit der Dimensions-Tabelle gebildet (Partial Partition Wise Join).

Ist die Dimensions-Tabelle ebenfalls und nach dem gleichen Partitionierungsschlüssel wie die Fakten-Tabellen partitioniert, so werden einzelne Teil-Joins zwischen den zusammengehörenden Partitionen gebildet (Full Partition Wise Join). Dieses Vorgehen findet auch bei Sub-Partitionierung statt.

Unterstützung im Ladeprozess

Partitioning bietet immer einen leichten Zugriff auf die Daten einer Ladeperiode. Wählt man für die neu zu ladenden Daten einer Ladeperiode eine Partition, so kann man diese schnell zu einer Warehouse-Tabelle hinzufügen oder im Fehlerfall wieder entfernen. Dieses Verfahren nennt sich „Partition Exchange and Load“ (PEaL). In fünf einfachen Schritten lässt sich so – ohne die vorhandene partitionierte Tabelle zu beinträchtigen und Abfragen darauf zu behindern – eine ebenfalls sehr große partitionierte Tabelle pflegen. Dazu hängt man zunächst eine leere Partition an die bestehende Tabelle an. Dann werden der Ladevorgang in einer separaten (temporären) Tabelle vollzogen und auch dort die Indizes gepflegt. Anschließend lässt sich in einem Schritt die leere Partition gegen die gepflegte Tabelle austauschen. Die neuen Daten stehen mit gepflegtem Index sofort zur Verfügung.

Diese Vorgehensweise vereinfacht viele ETL-Aufgaben und ermöglicht

auch Datenpflege während eines Online-Betriebs. Es ist eine hilfreiche Unterstützung, wenn man bei einem ETL-Schritt immer eine ganze Tabelle neu erzeugt, da dann die bestehende Tabelle mit einer Partition gebildet wird und die neu gepflegte separate, temporäre Tabelle in einem Schritt die bestehende ersetzt. Die Partitionierung bietet gerade in diesem Bereich eine enorme Hilfestellung, die zumindest die Wartungszeit verkürzt.

Vereinfachte Handhabung bei der Indizierung

Die Aktualisierung der Indizes stellt im ETL einen großen zeitlichen Aspekt dar. Partitionen hingegen, wie schon beim PEaL beschrieben, lassen sich separat voneinander indizieren (Local Index). Damit spart man sich das zeitaufwändige Aktualisieren eines kompletten (globalen) Index. Da sich Indizes ebenfalls partitionieren lassen, kann man auch einen Index anders partitionieren als die zugrunde liegende Tabelle. Dies ist in speziellen Situationen sinnvoll.

Vereinfachtes Update der Materialized View

Im Data Warehouse kommen oft Materialized Views (MAV) zum Einsatz. Diese führen meist als Aggregations-Tabellen zur Performance-Steigerung. Die Pflege dieser MAV kann durch Log-Tabellen an den zugrunde liegenden Tabellen oder aber über das Partition Change Tracking (PCT) erfolgen. Die Datenbank erkennt, welche Partitionen sich verändert haben oder neu erstellt wurden, und kann dann bei der Aktualisierung einer MAV selber entscheiden, welche Deltas nachzuladen sind. Dies ist in vielen Situationen schneller und effizienter als die MAV-Logs. Auch hier bietet das Partitionierungsverfahren einen erheblichen Vorteil beim Management des Data Warehouse.

Unterstützung bei der Komprimierung

Oracle bietet unterschiedliche Komprimierungsvarianten an. Partitioning kann je nach Bedarf Daten derselben Tabelle unterschiedlich komprimieren. OLTP-Komprimierung ist das kostenpflichtige Feature „Advanced Kompression“ der Datenbank. Es komprimiert Tabellen oder Partitionen, die noch geändert werden. Für Updates und Inserts bietet sich die kostenfreie Basic-Kompression nicht an. Durch Partitionierung ist es aber möglich, aktive Partitionen mit OLTP oder gar nicht zu komprimieren und inaktive Partitionen mit „Basic“. Dieser Mix spart Plattenplatz und durch die geringere Menge der zu entkomprimierenden Daten werden die Abfragen sogar schneller. Partitioning ist daher in den meisten Data-Warehouse-Systemen unter diesem Gesichtspunkt dringend zu empfehlen.

Unterstützung im Backup-Prozess

Die meisten Daten eines Data-Warehouse-Systems ändern sich nicht mehr, nachdem sie einmal gespeichert sind – und das oft über Jahre hinweg. Deswegen sind auch nur die immer wieder aktuell hinzukommenden Daten zu sichern. Große Tabellen lassen sich durch Partitioning leicht in Partitionen aufteilen, die sich geändert haben, und solche, die nur ältere Daten beinhalten. Nur die Änderungen sind beispielsweise mit RMAN zu sichern.

Durch die Möglichkeit, mehrere Partitionen zusammenzufügen, lassen sich Monatspartitionen des Vorjahres zu einer Vorjahres-Partition zusammenfassen und archivieren. Diese lassen sich auch von der Tabelle abschneiden und nur im Revisionsfall wieder anhängen. Die Archivierung wird damit sehr vereinfacht und ein Full-Backup, wie leider manchmal immer noch zu sehen, entfällt.

Unterstützung des Information Lifecycle Managements

Beim Information Lifecycle Management (ILM) speichert man unterschiedlich alte und unterschiedlich häufig genutzte Daten auf verschieden teure, schnelle und ausfallsichere Platten. Während traditionelle Speichermittel (Plattensysteme) immer günstiger geworden sind, ist ihre Performance gleich geblieben. SSD-Platten sind heute erschwinglich geworden und können partiell Spindel-Festplatten ersetzen. Sie sind allerdings immer noch zu teuer, um etwa 50 Terabyte große DWH-Systeme zu versorgen.

Die aktuellen Daten einer mit historischen Daten gefüllten Fakten-Tabelle hingegen können durchaus auf einzelnen SSD-Platten gelagert sein. Das Partitionieren ermöglicht dies durch das Zuweisen verschiedener Datenträger für die unterschiedlichen Partitionen. Beim ILM werden also verschiedene, bereits erwähnte Möglichkeiten einer Partitionierung ausgenutzt und zusammen angewendet. Der Enterprise Manager verwaltet die Partitionierung und steht zur Analyse zur Verfügung.

Fazit

Die Partitionierung bietet einen enormen Vorteil bei der Handhabung, Erstellung, Wartung und Abfrage von (nicht nur) großen Tabellen im Data Warehouse und ist in vielen Fällen ein Muss.

Detlef E. Schröder
detlef.e.schroeder@oracle.com

Alfred Schlaucher
alfred.schlaucher@oracle.com

Beim Aufbau von Data-Warehouse- und OLAP-Systemen hat Modellierung eine zentrale Bedeutung, da sie die Analyse- und Auswertungs-Möglichkeiten festlegt. Bereits auf der konzeptionellen Ebene sind die Anforderungen an Business-Intelligence-Systeme zu berücksichtigen. ADAPT, eine Modellierungssprache zur Definition multidimensionaler Datenstrukturen, wird von System-Architekten in der Praxis eingesetzt. Als Modell-Editor wird in der Regel Microsoft Visio mit speziellen ADAPT-Schablonen genutzt.

Vergleich zweier unterschiedlicher Ansätze zur Modellierung von OLAP-Systemen

Michael Weiler, PROMATIS software GmbH

Der Artikel vergleicht die ADAPT-Modelle mit einer objekt-relationalen Modellierung von Datenmodellen – exemplarisch dargestellt mit dem frei verfügbaren Werkzeug Horus. Kann ein Werkzeug, das seine Stärken in der kollaborativen Geschäftsprozessmodellierung besitzt, auch für die Modellierung von Business-Intelligence-Systemen genutzt werden? Beide Ansätze werden zunächst vorgestellt und danach unter verschiedenen Gesichtspunkten verglichen und bewertet.

ADAPT-Modellierung

Die Modellierungssprache ADAPT wurde im Jahr 1998 von Dan Bulos veröffentlicht und ist eine eingetragene Marke der Symmetrie Corporation. ADAPT steht für „Application Design for Analytical Processing Technologies" und soll die Unzulänglichkeiten klassischer Entity-Relationship-Modellierung beim Entwurf multidimensionaler Strukturen beheben. Hierzu existieren neun Grundelemente. Die beiden Kernelemente sind der „Cube" (Würfel) und die Dimension. Diese orientieren sich damit an den multidimensionalen Strukturen eines BI-Systems. Um die Hierarchien eines multidimensionalen Datenmodells abbilden zu können, gibt es die Elemente „Hierarchy" und „Level", wobei „Level" eine Ebene innerhalb einer Hierarchie bezeichnet. Zur weiteren Spezifikation der Dimensionen existieren die Elemente „Member", „Attribute" und „Scope". Als „Member" wird eine konkrete Ausprägung einer Dimension bezeichnet. Durch Attribute können weitere Informationen zu einer Dimension definiert werden. Über das Konstrukt „Scope" lässt sich eine Sammlung mehrerer „Members" gruppieren. Die letzten beiden Objekte sind das „Model" und der „Context". Eine beliebige Funktion zur Berechnung abgeleiteter Kennzahlen wird als „Model" bezeichnet. Mit dem „Context" kann ein Ausschnitt eines „Cube" modelliert werden. Abbildung 1 zeigt alle ADAPT-Objekte und deren Symbole.

Die einzelnen Objekte werden über unterschiedliche Verbindungen miteinander in Beziehungen gesetzt. Der einfache Pfeil beschreibt eine Verbindung vom übergeordneten zum darunterliegenden Objekt. Beispielsweise stellt man die Hierarchie und ein Attribut mit dem „einfachen Pfeil" von der Hierarchie zum Attribut dar. Die verschiedenen Ebenen einer Hierarchie werden über „Strict Precedence" oder „Loose Precedence" miteinander verbunden. Durch die Verbindung „Loose Precedence" lassen sich nicht balancierte Hierarchien modellieren. Die Pfeile vom Typ „Used By" stehen für Funktionen, um zu verdeutlichen, welche Attribute in der Berechnungsvorschrift genutzt werden. Es existieren

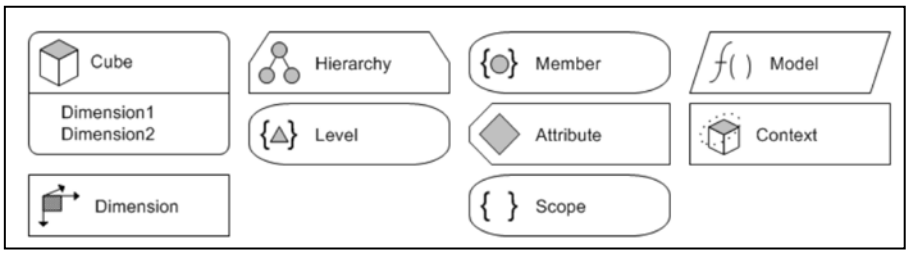


Abbildung 1: Die ADAPT-Symbole

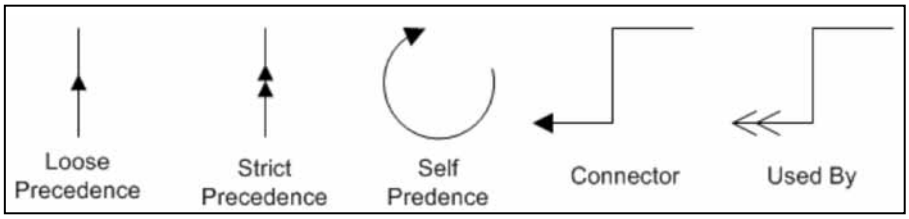


Abbildung 2: ADAPT-Verbindungselemente

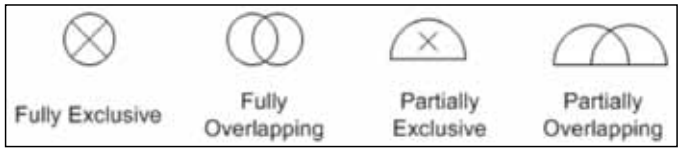


Abbildung 3: ADAPT-Operatoren (Abbildung 4 siehe www.doag.org/go/doagnews/weiler_abb4)

noch weitere Verbindungsarten, die jedoch in der Praxis kaum vorkommen. Abbildung 2 zeigt die wichtigsten Verbindungsarten.

Die Operatoren zum Bilden von Dimensions-Ausschnitten (Scope) stellen die letzte Elementklasse dar. Dabei wird beschrieben, ob Elemente vollständig oder teilweise exklusiv oder überlappend sind (siehe Abbildung 3).

ADAPT-Modell-Beispiel „Verkaufsanalyse"

Die multidimensionale Modellierung eines Würfels zur Analyse von Verkäufen zeigt beispielhaft die Nutzung der ADAPT-Elemente. Sehr häufig werden hierzu mehrere Zeichenblätter verwendet, um den Zusammenhang der Dimensionen und der Fakten zu definieren. Auf einem Übersichtsblatt befinden sich lediglich die Fakten und Dimensionen ohne Hierarchien, Levels etc. Auf den weiterführenden Blättern sind die Details zu einer Dimension dargestellt.

Im Beispiel der Verkaufsanalyse (siehe www.doag.org/go/doagnews/weiler_abb4) wurden alle vorgestellten Modellelemente eingesetzt. Das Modell beschreibt eine Verkaufsanalyse mit fünf Dimensionen. Es wurden lediglich die Dimensionen „Produkt", „Zeit" und „Metrik" ausmodelliert. Die Dimensionen „Kunde" und „Organisation" wurden nicht weiter verfeinert. Das Produkt besitzt zwei Attribute – einen deutschen und einen englischen Beschreibungstext. Selbstverständlich sind beliebige weitere Attribute möglich.

Die Produkt-Dimension hat zwei Hierarchien: eine für die Kategorien und eine für die Lieferanten. Enthält eine Hierarchie eine Gesamtsumme, so wird diese durch Nutzung der Verbindung „Strict Precedence" zum Hierarchie-Knoten dargestellt, wie dies im Beispiel bei der Lieferanten-Hierarchie modelliert wurde. Im Falle der Kategorie wurde bewusst darauf verzichtet. Dies bedeutet, die oberste Hierarchie-Ebene sind die einzelnen Kategorien. Die unterste Ebene beider Hierarchien sind die einzelnen Produkte. Jedes Produkt muss einem Lieferanten zugeordnet sein. Im Fall der Kategorien kann ein Produkt einer Unter-Kategorie oder einer Produkt-Kategorie zugeordnet sein. Dies wird durch den Verbindungstyp „Loose Precedence" vom Produkt zur Produkt-Unterkategorie definiert.

Ein Produkt muss eindeutig der Klasse „A" oder „B" zugeordnet sein. Produkte, die keiner Klasse zugeordnet sind, existieren nicht (Fully Exclusive). Ein Produkt ist im Katalog für das erste Halbjahr und oder im Katalog für das zweite Halbjahr enthalten. Produkte, die in keinem der beiden Kataloge enthalten sind, existieren nicht (Fully Overlapping). Es werden Produkte vom Typ „Verbinder" oder vom Typ „Befestigungen" verkauft. Dabei ist jedes Produkt eindeutig dem entsprechenden Typ zugeordnet. Zusätzlich existieren Produkte, die keinem der beiden Typen zugeordnet sind (Partially Exclusive). Die Produkte werden in Spezialkatalogen angeboten, wobei ein Produkt in beiden Katalogen vorkommen kann. In den Spezialkatalogen sind jedoch nicht alle Produkte enthalten (Partially Overlapping).

Bei der Kalender-Hierarchie wurde eine klassische Hierarchie mit „Monat", „Quartal" und „Jahr" dargestellt. Zudem bestehen auf Monatsebene mehrere Ausprägungen: „Aktueller Monat", „Vorheriger Monat" und „Monat Vorjahr". Ein Modell benutzt die beiden erstgenannten Ausprägungen und ermittelt eine Abweichung auf Basis des Umsatzes. Dies ist eine sehr genaue Definition eines Modells. Oftmals werden ADAPT-Modelle auf sehr hoher Ebene definiert, wie im Falle des Forecast-Modells, das auf Basis des Verkaufswürfels einen Forecast-Teilausschnitt berechnet. Eine detaillierte Berechnungsvorschrift ist oftmals in der Analyse eines BI-Systems nicht notwendig.

Die Horus-Methode

Als Ergebnis langjähriger Forschungsarbeit mit Universitäts- und Forschungsinstituten sowie mit einem Industriepartner ist unter dem Namen „Horus" eine völlig neue Generation von Tools zur Unterstützung des gesamten Lebenszyklus von Geschäftsprozessen entstanden. Mit dem Horus Business Modeler werden Geschäftsprozesse aus verschiedenen Blickwinkeln modelliert (Abläufe, Objekte, Organisationen, Ziele etc.). Das vereinfacht die einzelnen Modelle und erhöht die Flexibilität der Modellgestaltung. Aber: Oft überfordern viele Modelltypen den Nutzer. Deshalb werden lediglich vier unterschiedliche Modellierungssprachen verwendet. Neben Petri-Netzen (XML-Netze) für die Ablaufmodellierung kommen Organigramme, semantisch-hierarchische Strukturen und ein Objekt-Relationship-Modell zur Anwendung. Zur Modellierung von Business-Intelligence-Systemen haben sich Zielmodelle (modelliert als semantisch-hierarchische Struktur) zur Definition der Ziele an das neue System, Ablaufnetze zur Analyse der ETL-Prozesse und Objektmodelle für die Definition der multidimensionalen Strukturen bewährt. Im vorliegenden Fall wird die ADAPT-Modellierung mit dem Horus-Objektmodell verglichen.

Die nachfolgend beschriebene Notation zur Geschäftsobjekt-Modellierung wird auch kurz als „Objektmodell" bezeichnet. Für eine entsprechende Modellierung stehen die folgenden Grundelemente zur Verfügung:

- Objekt mit Attributen (einfache Geschäftsobjekt-Struktur)
- Aggregationstyp aus Objekten (komplexe Geschäftsobjekt-Struktur)
- Zwei Arten von Verbindungstypen zwischen einfachen Geschäftsobjekt-Strukturen (Objekte): Beziehungs- und Vererbungskanten
- Sammelbedingungen für Kanten

Ein „Objekt" ist ein Container, um die Attribute von Geschäftsobjekten in logischen Einheiten zusammenzufassen. Es besitzt einen eindeutigen Namen, optional ein oder mehrere Schlüsselattribute, weitere Attribute und gegebenenfalls Bedingungen. Kopie-Objekte stellen Referenzen auf andere Objekte dar, werden durch gestrichelte Rahmen dargestellt und können inhaltlich nicht geändert werden. Diese Kopien werden zur besseren Strukturierung der Modelle verwendet und hinter einer Objektkopie kann sich ein weiteres Teilmodell mit beliebig vielen Objekten verbergen (siehe Abbildung 5). Jedes Attribut ist mit einem Da-

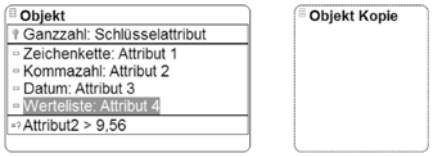


Abbildung 5: Horus-Objekt und Objektkopie

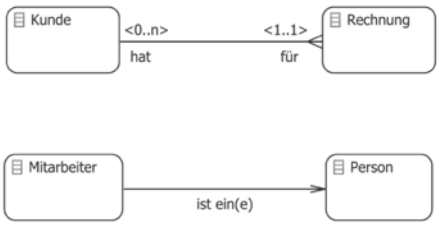


Abbildung 6: Horus-Verbindungstypen im Objektmodell

tentyp verbunden. Um Ausprägungen festhalten zu können, kann als spezieller Datentyp „Werteliste" verwendet werden, der eine Liste von „Ausprägungen" enthalten kann.

Verbindungstypen zwischen Objekten sind Elemente, die eine Klassifizierung gleichartiger Beziehungen oder Abhängigkeiten abbilden. Dabei kommen zum einen die aus der Entity-Relationship-Modellierung bekannten Beziehungstypen zum Einsatz und zum anderen eine Vererbungskante. Abbildung 6 zeigt ein Beispiel für beide Verbindungstypen.

Auf den Beziehungskanten können zusätzlich Sammelbedingungen genutzt werden. Eine XOR-Sammelbedingung ist ein „Ausschließendes Oder" der an die Sammelbedingung angeschlossenen Kanten. Eine OR-Sammelbedingung ist ein „Oder mit mindestens einer Auswahl".

Abbildung 7 zeigt jeweils ein Beispiel für eine XOR-Sammelbedingung bei Vererbungskanten und eine OR-Sammelbedingung bei Beziehungskanten. Im ersten Beispiel wird ausgedrückt, dass ein verkauftes Produkt entweder einen Produktionsauftrag oder einen Beschaffungsauftrag auslöst. Das zweite Beispiel zeigt, dass eine Bestellung genau eine Dienstleistung oder einen Artikel enthalten kann. Eine Bestellung kann im dargestellten Beispiel auch eine Dienstleistung und einen Artikel enthalten. Jedoch muss eine Bestellung mindestens eine Dienstleistung oder alternativ einen Artikel enthalten. Artikel und Dienstleistungen können jeweils in beliebig vielen Bestellungen enthalten sein. Es existieren noch weitere Konstrukte, die jedoch für das vorgestellte Beispielmodell nicht von Bedeutung sind.

Abbildung 7: Horus-Sammelbedingungen im Objektmodell

Horus-Objektmodell-Beispiel „Verkaufsanalyse"

Das bereits vorgestellte ADAPT-Modell wurde mit den Konstrukten des Horus-Objektmodells umgesetzt (siehe www.doag.org/go/doagnews/weiler_abb8). Obwohl das gesamte Modell auf einer Seite dargestellt ist, wurden bereits Objektkopien verwendet, über die, wie bereits erwähnt, die Detailinformationen zu einer Dimension in einem separaten Modell dargestellt werden können.

Die Formel für die Abweichung des Monats wurde über eine Bedingung eingegeben, die die konkrete Rechenvorschrift enthält. Ein einfaches Attribut mit der Rechenvorschrift in der Beschreibung stellt eine sinnvolle Alternative dar. Das Forecast-Modell wurde über die Vererbungskante modelliert. Eine detaillierte Beschreibung des Forecast-Modells kann zum einen in der Beschreibung oder durch die Verwendung eines Ablaufdiagramms erfolgen. Die ADAPT-Operatoren können entweder explizit über Objekte (Katalog 1. HJ, Katalog 2. HJ) oder einfacher über ein Attribut mit einer Werteliste (Attribut „Produktklasse") modelliert werden.

Vergleich beider Modellierungsvarianten

Um die Modelltypen miteinander zu vergleichen, wurden einige Kriterien definiert und für diese Kriterien Punkte zwischen „1" und „5" vergeben (siehe www.doag.org/go/doagnews/weiler_tabelle). Je mehr Punkte vergeben wurden, desto besser ist das Kriterium erfüllt. Bei einigen Kriterien wurde in Klammer eine Bewertung im Vergleich zur Horus Enterprise Version vorgenommen. Um eine Gesamtpunktzahl zu erhalten, sind die Kriterien unterschiedlich gewichtet.

Zusammenfassend zeigt Horus in der direkten Gegenüberstellung mehr Vorteile auf. Vergleicht man die Visio-Variante mit der Horus Enterprise Edition, so fällt der Unterschied deutlicher aus. Beide Techniken können zur Modellierung multidimensionaler Strukturen genutzt werden und an der einen oder anderen Stelle entscheiden die Vorlieben. Weitere Informationen zu ADAPT findet man auf www.symcorp.com und zu Horus unter www.horus.biz.

Fazit

Der Artikel stellt die ADAPT-Modelle und die Horus-Objektmodelle gegenüber. Beide Modellierungsarten haben Vor- und Nachteile. Die ADAPT-Modelle werden aufgrund der expliziten Darstellung von Attributen und Gruppen rasch sehr umfangreich. Die Horus-Objektmodelle sind kompakter. Durch die vielen verschiedenen Modellierungselemente bei ADAPT kann ein genaueres Modell erstellt werden. Ein Architekt muss sich jedoch fragen, ob die Modelle eine verbesserte Aussagekraft besitzen. Die vordefinierten Felder in Horus ermöglichen die genaue Spezifikation der einzelnen Objekte und Attribute. Bei den Schablonen für Microsoft Visio ist dies nicht enthalten und selbst zu definieren. Die Verbindung von Objekt- und Ablaufmodellen sowie der Einsatz weiterer Modell-

arten zur ganzheitlichen Definition eines BI-Systems sprechen eindeutig für die Horus-Methode. Mit Visio kann dies ebenfalls erreicht werden, fordert aber einen klaren Style-Guide, welche Objekte genutzt und wie diese miteinander verbunden werden.

Bei der Vollversion von Horus sind die Generierung einer Gesamt-Dokumentation und damit die vollautomatische Pflichtenheft-Erstellung sowie eine Bereitstellung der Modelle in einem unternehmensinternen Wiki weitere Pluspunkte. Die automatische Erstellung von Skripten für unterschiedliche ETL-Werkzeuge ist derzeit nur angekündigt und würde den Phasenübergang von der Konzeption und vom Design in die Implementierung erheblich erleichtern. Insgesamt geht Horus als Punktsieger aus dem Vergleich hervor, wenn auch mit der ADAPT-Methode eine sinnvolle Modellierung für multidimensionale Strukturen möglich ist. Wichtig ist, dass überhaupt eine Konzeptionsphase mit Modellen unterstützt wird, um den optimalen Nutzen des zukünftigen BI-Systems innerhalb eines festen Zeitrahmens und Budgets zu erreichen.

Michael Weiler
michael.weiler@promatis.de

Aus dem Verein

Stefan Kinnen
Leiter der Development Community
dec@doag.org

Neues aus der Development Community

Das war sie also: die Feuertaufe für eine eigene Development-Fachkonferenz. Mehr als 200 Teilnehmer folgten der Einladung am 14. Juni 2012 nach Bonn und übertrafen damit die Erwartungen der DOAG deutlich. Bei der Zusammensetzung des Programms mit vielen namhaften Referenten war das auch verständlich.

Wussten Sie beispielsweise, dass es mittlerweile elf Programmiersprachen gibt, mit denen Oracle-Programme geschrieben werden können? Oder dass es gegenüber rund 75.000 PC-Programmen etwa 1.3 Millionen Handy-Apps gibt? Daniel Liebhart spannte diesen Bogen über seine absolut kurzweilige Keynote und zeigte auf, dass wir nach der Ära „Host-Computing und Client/Server" nun in der dritten Generation von Software-Architekturen angekommen sind.

Datenbanknah prägt natürlich Apex das Geschehen. Mit der Version 4.2 kommen wieder wertvolle neue Features, die Apex beispielsweise noch deutlich weiter für Mobile Computing vorbereiten. Genauso wichtig in der DOAG-Community sind aber auch die immer zahlreicher werdenden Praxisberichte, die darstellen, was mit Apex heute bereits wirklich produktiv nutzbar umgesetzt werden kann.

Als breite Basis der Anwendungsentwicklung steht natürlich noch immer Java im Mittelpunkt. Aus Sicht der Oracle-Anwender kommen immer wieder Fragen nach Möglichkeiten und Erfahrungen der Migration von Forms-Applikationen in Richtung „Java" auf. Neue Tool-Unterstützung und spezielle Frameworks fanden zu Recht viel Aufmerksamkeit in Bonn.

Bei den eigentlichen Entwicklungs-Werkzeugen reichte ein eintägiger Stream wirklich nur dazu aus, um einige punktuelle Einblicke – beispielsweise in die New Features des JDeveloper 12c – zu geben. Ob und wie der BI Publisher als Reporting-Tool eingesetzt werden kann, wurde bereits am Vorabend im Rahmen eines Regionaltreffens NRW live präsentiert. Zu guter Letzt muss sich auch der konsequenteste Datenbank-Anhänger irgendwann mit NoSQL und somit quasi folgerichtig mit Big Data beschäftigen – so geschehen im vierten Stream „BPM und Software-Architektur". Das Fazit der DOAG 2012 Development: Prima, weiter so! Somit laufen bereits die Planungen für eine Wiederholung im nächsten Jahr. Dort werden wir das Motto „Software-Entwicklung auf Basis von Oracle-Tools und -Technologien – wohin die Reise geht" bestimmt weiter vertiefen können.

Dr. Frank Schönthaler
Leiter der Business Solutions Community
frank.schoenthaler@doag.org

DOAG 2012 Applications Konferenz + Ausstellung: Business Excellence im Visier

Die Business Solutions Community der DOAG traf sich vom 8. bis 10. Mai 2012

zu ihrem Top-Event im Herzen Berlins. Die DOAG 2012 Applications Konferenz + Ausstellung ist Europas führende Konferenz rund um Geschäftsprozesse, Oracle-Business-Applikationen und die zugrunde liegenden Technologien. In hochkarätigen Keynotes, praxisnahen Fachvorträgen und der begleitenden Ausstellung standen „Geschwindigkeit, Sicherheit und Innovation mit Oracle-Applikationen“ im Fokus. Insbesondere am Workshop-Tag konnten sich Anwender und Experten direkt an der Quelle informieren und erhielten ausreichend Möglichkeiten zum Networking sowie zum Erfahrungsaustausch.

Dr. Frank Schönthaler, Leiter der DOAG BSC, eröffnete die Konferenz mit der Keynote zum Thema „Business Excellence in volatilen Märkten“. Die Wichtigkeit dieser Thematik bestätigte sich auch in der anschließenden Keynote von Christian Stengel, Oracle EMEA, der in seinem „Oracle Fusion Applications Update“ immer wieder Bezug auf die einleitenden Grundgedanken nahm. Anschließend folgten hochinteressante Vorträge in mehreren parallelen Streams.

Ein Highlight am Ende des zweiten Konferenztages war die Panel-Diskussion zur Entwicklung des deutschsprachigen Markts für Oracle-Geschäfts-Applikationen. Gerade von Seiten der Anwender und Implementierungspartner wurde die einseitige Fokussierung der Oracle-Marketingstrategie auf konkrete Demand-Generation-Aktionen kritisiert. Diese stößt teilweise auf Unverständnis, da sie vergisst, dass man zunächst Wahrnehmung für die Oracle-Applikations-Produkte im Markt schaffen muss. Dabei entflammte eine lebhafte Diskussion, die dem Partner Oracle in gewisser Weise einen Meinungsquerschnitt über die Applikations-Kundenbasis im deutschsprachigen Raum brachte.

Die Workshops am dritten Konferenztag boten einen sehr großen Mehrwert sowohl für die Teilnehmer als auch für die Referenten. Es wurden weniger Produkt-Themen, als vielmehr konkrete Business-Themen diskutiert und gezeigt, wie diese mit Oracle-Produkten abgebildet werden können. BSC-Leiter Dr. Frank Schönthaler blickt mit allen Community-Leitern hochzufrieden auf die Veranstaltung zurück: „Dass die Teilnehmerzahl der DOAG 2012 Applications gegenüber dem Vorjahr noch einmal gesteigert werden konnte, hebt die Bedeutung dieser Veranstaltung im deutschsprachigen Raum hervor.“

Integrata-Kongress 2012 – Mehr Lebensqualität durch IT!

Vom 10. bis 11. Mai 2012 fand im Anschluss an die DOAG 2012 Applications Konferenz + Ausstellung der 2. Kongress der Integrata-Stiftung in Berlin statt. Im Fokus der Stiftung steht die „Humane Nutzung der Informationstechnologie – Mehr Lebensqualität durch IT!“ Die DOAG eröffnete ihren Mitgliedern dieses zusätzliche Angebot, da auch sie erkennt, dass Nachhaltigkeit und Lebensqualität vor allem langfristig immer mehr in den gesellschaftlichen Mittelpunkt rücken. Der Premium-Sponsor Oracle unterstrich durch sein Engagement auf dem Kongress die Wichtigkeit dieser Thematik für die Entwicklung von Oracle und auch für deren IT Produkte.

Unter dem diesjährigen Konferenz-Motto „Mehr Demokratie durch IT!“ etablierte sich die Integrata-Stiftung als eine lebendige Plattform zum Mitdenken. Julian Nida-Rümelin, Staatsminister a.D. und Präsident der Deutschen Gesellschaft für Philosophie, eröffnete den ersten Konferenztag mit einer starken Keynote zum Konferenz-Motto. Anschließend folgte der Honorar-Professor für Wirtschaftsethik und Präsident des Niedersächsischen Landesamts für Lehrerbildung und Schulentwicklung a.D., Prof. Wolf Dieter Hasenclever. In seinem Beitrag „Bildung und eCommunication – wie sich das Lernen verändert“ setzte er sich kritisch mit den Veränderungen des Lernverhaltens bei Jugendlichen auseinander, die vor allem durch die Verbreitung von Smartphones getrieben sind. Er sieht die IT und deren Möglichkeiten aber auch als einzige Chance, um globalen Wissenstransfer langfristig gewährleisten zu können. Franz Reinhard Habbel, Sprecher des Deutschen Städte- und Gemeindebundes (DStGB), beendete den ersten Konferenztag mit dem Thema „Zeitwende – Politik 2012“. Er zeigte auf, wie eGovernance heute funktioniert und versprach: „Transparenz, Partizipation und Offenheit gewinnen an Bedeutung“.

Den zweiten Konferenztag leitete Christof Leng, Vizepräsident der Gesellschaft für Informatik und Mitgründer der Piratenpartei, mit der Keynote „Der Arabische Frühling: Gefahren und Chancen der IT für die Demokratisierung“ ein. Hierbei thematisierte er die bedeutenden Umwälzungen in Nordafrika, die oft zweischneidige Rolle der Kommunikations- und Informationstechnologie und die bisher nicht abschätzbaren Folgen der Nutzung von IT in diesen Regionen.

Es folgten drei hochwertige, parallele Streams. Michael Mörike, Vorstand der Integrata-Stiftung, moderierte den Stream „Politische Partizipation“, Welf Schröter vom Forum Soziale Technikgestaltung den Stream „Befriedigende Arbeit“ und Prof. Dr. Marco Mevius von der HTWG Konstanz den Stream zum Thema „Gesunde Umwelt“. Alle Konferenzbeiträge erfuhren eine positive Resonanz.

Die begleitende Ausstellung lud die Besucher an beiden Konferenztagen zum ausgiebigen Informieren und „Networken“ ein und es wurden viele interessante Gespräche rund um Nachhaltigkeit und mehr Lebensqualität durch die Informationstechnologie geführt. Weitere Informationen zur Integrata-Stiftung unter http://www.integrata-stiftung.de.

Vorschau auf die nächste Ausgabe

Die Ausgabe 05/2012 hat das Schwerpunktthema

Middleware

Sie erscheint am 5. Oktober 2012

Christian Trieb
Leiter der Datenbank Community
und internationale Aktivitäten
dbc@doag.org

Aktuelles von der Datenbank Community

Die von der Datenbank Community jeden zweiten Freitag im Monat angebotenen Webinare werden so gut angenommen, dass die Lizenz zur Durchführung von 25 auf 100 Teilnehmer erweitert werden musste. Das Feedback der Teilnehmer ist sehr positiv. Wer interessiert ist, ein Webinar zu einem Datenbank-Thema zu halten, oder Themenwünsche hat, kann sich gerne an die Community-Leitung unter dbc@doag.org wenden.

Die Datenbank Community begrüßt herzlich Tilo Metzger von der Firma M-experience Multimedia, der zukünftig als zusätzlicher Leiter der SIG Security Franz Hüll unterstützen wird. Ebenso herzlich heißen wir Andreas Buckenhofer von der Firma Daimler TSS GmbH willkommen, der in der Datenbank Community aktiv mitarbeiten wird.

Für den 15. Mai 2013 plant die DOAG eine Datenbank-Fachkonferenz in Düsseldorf. Dort sollen innerhalb eines Tages die unterschiedlichen Aspekte der Oracle-Datenbank präsentiert und intensiv diskutiert werden. Auch eine begleitende Fachausstellung ist auf dem Plan.

Höhepunkt der Datenbank Community in diesem Jahr ist die DOAG 2012 Konferenz + Ausstellung, die vom 20. bis 22. November 2012 in Nürnberg stattfindet. Am ersten Tag wird es abends ein informelles Treffen der Datenbank-Interessierten geben. Während der dreitägigen Konferenz stehen rund 100 sehr interessante Vorträge zu allen Datenbank-Themen auf dem Programm. Hinzu kommen zwei Experten-Panels. In Zusammenarbeit mit der SIG RAC ist zum ersten Mal ein RAC-Attack-Workshop geplant.

Berliner Expertenseminar „Partitionierung“

Am 10. und 11. Juli 2012 fand in der DOAG-Konferenz-Lounge das Berliner Expertenseminar zum Thema „Oracle Partitionierung“ statt. Referent war Klaus Reimers von der ORDIX AG. Er ging während des Seminars auf alle Ausprägungen der Partitionierung in einer Oracle-Datenbank ein und stellte sie anschaulich und anhand von Kundenbeispielen sehr gut dar. Die Konzepte und Techniken wurden präsentiert sowie deren sinnvolle Verwendung erläutert.

Die Teilnehmer wurden gut in das Seminar einbezogen und es gab immer die Möglichkeit des Nachfragens, die auch intensiv genutzt wurde. Dies macht einen der entscheidenden Vorteile der DOAG-Expertenseminare deutlich. Durch die kleine Seminargruppe, einen kompetenten Referenten und ein fokussiertes Thema gelingt es sehr gut, in einen konstruktiven Dialog zu treten, der es ermöglicht, das Vorgetragene auch gut verstehen. Dazu tragen natürlich auch das Ambiente der DOAG KonferenzLounge und die gelungene Abendveranstaltung bei.

Wer Themenwünsche oder Themenvorschläge hat, kann sich gerne an die DOAG unter expertenseminare@doag.org wenden.

Internationales

Am 22. und 23. Mai 2012 trafen sich Vertreter der Oracle-Anwendergruppen aus der EMEA-Region zum jährlichen Austausch in Riga. Die DOAG war durch den Vorstand Ralf Kölling und Christian Trieb, Leiter der Datenbank Community, vertreten. Fried Saacke, Geschäftsführer der DOAG Dienstleistungen GmbH, vertrat den Interessenverbund der Java User Groups e.V. (iJUG), bei dem die DOAG Mitglied ist.

Insgesamt waren rund 40 Vertreter der verschiedensten Anwendergruppen anwesend, sodass sich schnell eine gute und abwechslungsreiche Kommunikation entwickelte. Die Themen waren unter anderem der Austausch von Referenten für Anwendergruppen-Veranstaltungen, Finanzierungsmöglichkeiten für Anwendergruppen, Zusammenarbeit mit dem Oracle-Support sowie die Vorbereitung der Präsenz der EMEA-Anwendergruppen während der Oracle OpenWorld 2012 in San Francisco.

Seitens Oracle wurden die Möglichkeiten der Zusammenarbeit dargestellt. Hierbei lag der Schwerpunkt auf der sinnvollen Nutzung von Web-2.0-Techniken und den sich daraus ergebenden Vorteilen für die Anwendergruppen. Darüber hinaus zeigte Oracle die aktuellsten Entwicklungen der unterschiedlichen Produktgruppen auf.

Newsticker

Umfrage zur Virtualisierungs-Technologie
An der Umfrage der DOAG nahmen 212 Unternehmen teil, 89 Prozent davon haben bereits Virtualisierungs-Technologien im Einsatz. Die Vorteile der Virtualisierung liegen für die Befragten in den Einsparungen von Hard- und Softwarekosten (83 Prozent), in der besseren Systemauslastung (80 Prozent) und in den damit verbundenen Kosten-Einsparungen, etwa bei der Lizenzierung oder Administration der Systeme (70 Prozent).

62 Prozent der Unternehmen setzen VMware ESX Server ein, Oracle VM für x86 kommt auf 22 Prozent. Interessant ist das Zusammenspiel zwischen VMware und Oracle: 71 Prozent wünschen sich eine Änderung der Oracle-Lizenzpolitik im Zusammenhang mit VMware, 56 Prozent erwarten eine bessere Unterstützung von VMware unter Oracle. Die Bilanz von Oracle VM ist durchwachsen. Knapp die Hälfte (48 Prozent) derer, die das Produkt aktiv einsetzen, ist damit zufrieden bis sehr zufrieden; 43 Prozent sind unzufrieden bis sehr unzufrieden. Die Oracle-VM-Nutzer erwarten eine schnellere Weiterentwicklung (26 Prozent) und eine allgemein bessere Qualität des Produkts.

23.08.2012
Regionaltreffen NRW
Best Practices in DBnaher Programmierung
Stefan Kinnen
regio-nrw@doag.org

29.08.2012
SIG Middleware
Oracle Middleware,
Administration & Monitoring
Jan-Peter Timmermann
sig-middleware@doag.org

04.09.2012
SIG MySQL
MySQL Replikation
Matthias Jung
sig-mysql@doag.org

05.09.2012
Regionaltreffen Berlin/Brandenburg
Grid Control 11*g*
Michel Keemers
regio-bb@doag.org

11.09.2012
Regionaltreffen Hamburg/Nord
Stefan Thielebein
regio-nord@doag.org

11.09.2012
Regionaltreffen Jena/Thüringen
Jörg Hildebrandt
regio-thueringen@doag.org

12.09.2012
Regionaltreffen Rhein-Main
Thomas Tretter, Kathleen Hock
regio-rhein-main@doag.org

13.09.2012
Regionaltreffen Trier/ Saarland/Luxemburg
Oracle Advanced Queuing
Bernd Tuba
regio-trier@doag.org

13.09.2012
Regionaltreffen Würzburg
Shareplex Migration
Oliver Pyka
Regio-wuerzburg@doag.org

14.09.2012
Webinar
Flashback früher war alles besser
DOAG Geschäftsstelle
office@doag.org

18.09.2012/19.09.2012
Berliner Expertenseminar
„Performance“ mit Lutz Fröhlich
Cornel Albert
expertenseminare@doag.org

20.09.2012
Regionaltreffen Nürnberg/Franken
André Sept
regio-franken@doag.org

20.09.2012
SIG Database
Monitoring/Tools
Christian Trieb, Frank Stöcker
sig-database@doag.org

24.09.2012
Nordlichtertreffen der Regionalgruppen Bremen, Hamburg und Hannover
Ralf Kölling
regio-bremen@doag.org

24.09.2012
Regionaltreffen Osnabrück/ Bielefeld/Münster
Andreas Kother, Klaus Günther
regio-osnabrueck@doag.org

25.09.2012
Regionaltreffen NRW
Stefan Kinnen, Andreas Stephan
regio-nrw@doag.org

26.09.2012/27.09.2012
Berliner Expertenseminar
„Oracle Security“ mit Pete Finnigan
Cornel Albert
expertenseminare@doag.org

26.09.2012
SIG Development
APEX und Cloud Computing
Andreas Badelt, Christian Schwitalla
sig-development@doag.org

27.09.2012
Regionaltreffen München/Südbayern
Andreas Ströbel, Franz Hüll
regio-muenchen@doag.org

27.09.2012
Regionaltreffen Rhein-Neckar
Franz Stöcker
regio-rhein-neckar@doag.org

2012
DOAG
Konferenz + Ausstellung

20.11.2012 – 22.11.2012
DOAG 2012 Konferenz + Ausstellung
inkl. Schulungstag am 23.11.2012

Wir sind die Oracle-Community – unter diesem Motto kommen die Anwender aller Oracle-Produkte seit 25 Jahren zur jährlichen Anwenderkonferenz zusammen. Sie erhalten drei Tage Wissen pur, neueste Informationen zum erfolgreichen Einsatz der Oracle-Lösungen und praxisnahen Erfahrungsaustausch.

Den Teilnehmern eröffnet sich die attraktive Gelegenheit, ihr Netzwerk zu erweitern und von den Erfahrungen und dem Know-how der Oracle-Community zu profitieren.

http://2012.doag.org

heise Events
Conferences, Seminars, Workshops

in Kooperation mit DOAG Deutsche ORACLE-Anwendergruppe e.V.

Programm **Herbst 2012**

Konferenzen

RECHENZENTREN UND INFRASTRUKTUR 2012

Komponenten, Kabel, Netzwerke

- 27. September, Mannheim
- 28. November, Köln
- 11. Dezember, Hamburg

Android, iOS und Co. im Enterprise-Umfeld

Professioneller Umgang mit Smartphones und Tablets

- 5. September, Düsseldorf
- 11. September, Hamburg
- 13. September, München
- 19. September, Frankfurt

Cloud im Business-Einsatz

Einführung und Infrastruktur unter Berücksichtigung von Compliance Aspekten

- 18. September, München
- 20. September, Berlin
- 25. September, Köln

BYOD

heise Netze

Fremde Geräte im Netz

- 07. November, Hamburg
- 13. November, München
- 15. November, Stuttgart
- 29. November, Köln

Workshops / Seminare

SAN-Management: VMware ESXi und iSCSI

Speicherverwaltung mit Virtualisierungsservern

- 30. Oktober, Marburg

Kerberos – LDAP – Active Directory

Kerberos – Single Sign-On im gemischten Linux- und Windows-Umfeld

- 25. - 26. September, Köln
- 7. - 8. November, Hamburg
- 14. - 15. November, Stuttgart
- 4. - 5. Dezember, Hannover

Weitere Infos und Anmeldung unter: **www.heise.de/events**

Die Trivadis Innovation in Business Intelligence

trivadis
makes IT easier.

BASEL BERN LAUSANNE ZÜRICH DÜSSELDORF FRANKFURT A. M. FREIBURG I. BR. HAMBURG MÜNCHEN STUTTGART WIEN

biGenius™: DIE SUPERSCHLAUE DATA WAREHOUSE-LÖSUNG

Mit dem Trivadis Tool biGenius™ beschleunigen Sie die Implementierung Ihrer BI-Lösung. Sie werden von der Anforderungsanalyse bis zum automatisierten Generieren der Data Warehouse-Komponenten optimal unterstützt. Nutzen Sie die Vorteile des Trivadis Tools biGenius™ auch für Ihre Data Warehouse-Lösung und freuen Sie sich auf minimale Entwicklungskosten und hohe Flexibilität. Genial einfach eben.

Jetzt mehr erfahren unter www.trivadis.com/biGenius