Benchmarks und Standards

Ausarbeitung

zum Vortrag im Rahmen des Seminars

„Business Intelligence I: OLAP und Datawarehousing“

von Karl-Christian Pammer

email: pammer@rhrk.uni-kl.de

Betreuer: Dr. Ulrich Marder

18. Juli 2003

Inhaltsverzeichnis

1 Einleitung

Modelle zur Messung der Güte bzw. der Leistungsfähigkeit eine Systems, einer Maschine o. ä. haben in vielen Bereichen eine lange Tradition; als Beispiel lässt sich hier die Angabe zum Kraftstoffverbrauch eines Autos nennen. Dieses Beispiel zeigt jedoch, dass das gewonnene Ergebnis von vielen Faktoren (Fahrstrecke, Fahrstil etc.) abhängig ist. Aus diesem Grunde wurde ein vorgeschriebener Streckenmix (Stadtverkehr, Autobahn usw.) definiert, um die Angaben verschiedener Hersteller vergleichbar zu machen. Es wurden also Randbedingungen für die Testumgebung und den eigentlichen Test festgelegt.
Dieser Aspekt wurde umso wichtiger, da die Ergebnisse von Benchmarks intensiv zu Marketingzwecken – um die Überlegenheit des eigenen Produkts über die Mitbewerber herauszustellen – genutzt werden. Dieses sogenannte „Benchmarketing“ ist auch in der Informatik sehr verbreitet, etwa in Form von Integer- und Gleitkommaleistungsangaben bei Prozessoren. Solche Angaben haben jedoch nur eine geringe Aussagekraft bezogen auf die Gesamtleistung des Systems, da hier das Zusammenspiel der einzelnen Komponenten einen entscheidenen Einfluss hat.
Im folgenden werden zwei Benchmarks (TPC-D und APB-1) exemplarisch vorgestellt. Kern dieser Benchmarks sind Anwendungen aus dem Bereich OLAP und Decision Support. Die Datenbasis solcher Anwendungen ergibt sich häufig aus der Integration von Daten aus diversen Datenquellen. Dabei kann es sich um verschiedene (oft heterogene) Datenbanksysteme oder auch um verschiedenartige Datenquellen handeln. Aus diesem Grund werden im zweiten Abschnitt verschiedene Standards zur Datenintegration vorgestellt.

2 Benchmarks

Über einen langen Zeitraum existierten nur herstellerspezifische Benchmarks (z.B. TP1 von IBM), darum waren die Vergleichsmöglichkeiten von Systemen untereinander sehr beschränkt. Außerdem liegt der Verdacht nahe, dass der Benchmark sehr passgenau auf das zu testende System zugeschnitten wurde, da beide von dem selben Hersteller stammten.
Als Konsequenz aus dieser Situation begannen Organisationen wie etwa das

Transaction Processing Council (TPC-A, B, C, D, H, R und W) oder das OLAP Council (APB-1) Benchmarks zu spezifizieren. Ziel war es, einen Mechanismus zur Verfügung zu stellen, der Systeme über Hersteller- und Hardwaregrenzen hinweg vergleichbarer macht (siehe dazu [Burg00]).
In Benchmarks versucht man die Realität dadurch anzunähern, dass die in den Testläufen durchgeführten Operationen in ähnlicher Form auch in der Praxis auftreten. Da die Benchmarkgestaltung jedoch nicht auf ein bestimmtes Testfeld oder eine bestimmte Branche zugeschnitten ist, kann ein Benchmark eine detaillierte Systemevaluation – bezogen auf das Anwendungsszenario der späteren Nutzung – nicht ersetzen. Dies, so stellen beide Gremien in den Präambeln ihrer Benchmarks ausdrücklich klar, ist auch nicht ihr Ziel. Ein Benchmarkergebnis kann nur den Ausgangspunkt für die Evaluation von Systemen bilden.
Die Benchmark Spezifikation enthält neben einer Beschreibung der eigentlichen Aufgaben, die während eines Benchmarkdurchlaufes durchgeführt werden, zusätzlich weitreichende Anforderungen an den Benchmarknutzer und die Testumgebung.
Dies sind zum einen funktionale Aspekte für die Benchmarkdurchführung wie etwa kommerzielle Verfügbarkeit der verwendeten Software, Transaktionseigenschaften oder die Population der Testdatenbank.
Zum anderen definieren die Spezifikationen umfangreiche Informationspflichten bei der Veröffentlichung von Benchmarkergebnissen. So schreibt etwa der APB-1-Benchmark [OLAP98] vor, dass die Testumgebung so detailliert dokumentiert werden muss, dass sie rekonstruierbar ist und die erzielten Ergebnisse reproduziert werden können.
Darüber hinaus enthalten die Spezifikationen die Anforderung, dass die Ergebnisse durch einen Auditor verifiziert werden müssen. Dieser untersucht, ob die definierten Randbedingungen eingehalten wurden und besonders achten sie darauf, ob das System Funktionen enthält und nutzt, die speziell auf den Benchmark zugeschnitten sind. So zog etwa Oracle 1994 Ergebnisse des TPC-A Benchmarks zurück, weil ein speziell auf den Benchmark ausgerichteter Transaktions-Typ bei der Durchführung genutzt worden war [Shan98].

2.1 OLAP Council APB-1

Der APB-1 Benchmark [OLAP98] simuliert eine OLAP-Anwendung zur Analyse der Absatz- und Vertriebsdaten eines Unternehmens. Die verwendeten Analysemethoden sind nicht auf ein bestimmtes Geschäftsfeld oder eine spezielle Branche abgestimmt, vielmehr handelt es sich um eine Sammlung von allgemein akzeptierten Analysen, sogenannten „Best Practices“. Das Ziel von APB-1 ist es, die Gesamtleistung eines Serversystems bei OLAP-Anwendungen zu bestimmen; darum müssen alle Berechnungen und die gesamte Datenhaltung serverseitig erfolgen.
Die Datenbasis der Analyse besteht aus drei hierarchisch strukturierten Objekttypen, wobei die oberste Hierachiestufe („Top“) stets nur ein Element enthält:

- Channel
  Dieser Objekttyp enthält verschiedene Absatzwege, über die die Produkte vertrieben werden. Er besteht aus zwei Ebenen („Top“ und „Base“). Die Mindestanzahl der vorhandenen Objekte dieses Typs ist 10.

- Customer
  Dieser aus drei Ebenen („Top“, „Retailer“ und „Store“) bestehende Objekttyp enthält die Kunden, sowie die an die Kunden abgesetzten Mengen. Die Zahlen gliedern sich zum einen nach den verschiedenen Zwischenhändlern und für den jeweiligen Zwischenhändler noch einmal feingranularer auf dessen Geschäfte.
  Von diesem Objekttyp müssen 100-mal so viele Elemente existieren wie vom Typ „Channel“, also mindestens 1.000, außerdem müssen 90 Prozent der Daten zur untersten Hierachiestufe („Store“) gehören.

- Product
  Dieser Objekttyp besteht aus sieben Stufen („Top“, „Division“, „Line“, „Family“, „Group“, „Class“ und „Code“). Er beschreibt die Einbettung eines Artikels, bezeichnet durch seine Artikelnummer („Code“), in das Produkt-Sortiment oder -Portfolio.
  Die Kardinalität beträgt 10-mal die Kardinalität von „Customer“, mindestens jedoch 10.000 Objekte.

Um bei verschiedenen Benchmarkdurchläufen eine einheitliche Datenbasis zu garantieren – denn nur unter dieser Voraussetzung sind die Ergebnisse verschiedener Durchläufe vergleichbar – wird das Programm APB1GEN mitgeliefert. APB1GEN erzeugt die Objekte o. g. Typen in Form von nicht sortierten ASCII-Dateien, die dann in das OLAP-System eingelesen werden.
Zusätzlich zu den o. g. Typen definiert die Spezifikation noch zwei weitere, die zur Abbildung der zeitlichen Dimension der Analyse benötigt werden:

- Time
  Die Zeit-Dimension besteht aus einem 2-Jahreszeitraum (1995 und 1996). Sie enthält die Bestände per Monat und zusätzliche Abschlüsse per Quartal, per Jahr und von 1995 bis Simulationszeitpunkt Juni 1996 („Year-to-Date-Abschluss“). Durch die Festlegung des Simulationszeitpunktes Juni 1996 unterteilt sich die Zeit-Dimension in Vergangenheit (Januar 1995 bis Mai 1996) und Zukunft (Juli 1996 bis Dezember 1996).

- Scenario
  Ausgangspunkt der Szenarienanalyse sind zum einen der Year-To-Date-Abschluss („Actuals“) und zum anderen das Budget 1996. Zusätzlich erfolgt die Berechnung einer Vorhersage der Absatzentwicklung für das laufende Jahr aus den Daten das Vorjahres („Forecast“).
  Die Szenarioanalyse stellt dann das Budget oder die Vorhersage dem Abschluss per Juni 1996 gegenüber („Budget vs. Actuals“ oder „Forecast vs. Actuals“).

APB-1 betrachtet bei der Analyse zehn verschiedene Maßzahlen, die sich in zwei Klassen unterteilen lassen. Eine Klasse bilden die Maßzahlen, die sich direkt aus der Datenbasis bestimmen lassen („Input Measures“), dazu zählen beispielsweise die Absatzmenge. Zu diesen Maßzahlen existieren verschiedene Einflussfaktoren, die den Wert bestimmen. So ist etwa die Absatzmenge davon abhängig, ob sie über alle Kunden ober nur über einige berechnet wird.
Die folgende Tabelle zeigt die Abhängigkeiten:

Input Measures						
Maßzahl		Product	Customer	Channel	Time	Scenario
Absatzmenge	Units Sold	X	X	X	X	X
Umsatz	Dollar Sales	X	X	X	X	X
Bestand	Inventory	X	X		X	
Stückkosten	Product Costs		X		X	X
Versandkosten	Shipping Costs			X	X	X

Die zweite Klasse bilden die Maßzahlen, die aus denen der ersten Klasse berechnet werden („Calculated Measures“):

Calculated Measures		
Maßzahl		Berechnung
Durchschnittspreis	Average Price	$\frac{Umsatz}{Absatzmenge}$
Gesamtkosten	Cost	$Absatzmenge * (Stückkosten + Versandkosten)$
Marge	Margin	$Umsatz - Gesamtkosten$
Margenanteil	Margin Percent	$\frac{Marge}{Umsatz}$
Umsatzschnitt	Smoothed Sales	$Durchschnitt der letzten 6 Monate$

Diese Maßzahlen finden in den Abfragen des APB-1 Benchmarks Verwendung. Die Zusammenstellung der Abfragen („Query Mix“) besteht aus zehn verschiedenen Abfragen mit unterschiedlichen Ausführungshäufigkeiten (siehe Tabelle).

1. Channel Sales Analysis
   Bestimmt die Absatzmenge, den Umsatz und den Durchschnittspreis für einen Absatzkanal bis zum angegebenen Zeitpunkt.

2. Customer Margin Analysis
   Bestimmt die Umsätze, die Kosten und die Margen für einen spezifischen Kunden in einem vorgegebenen Zeitraum über alle Vertriebswege.

3. Product Inventory Analysis
   Bestimmt die Absatzmenge, den Umsatz, die Kosten und den Bestand für ein Produkt per Juni 1996.

4. Time Series Analysis

Bestimmt die Umsätze eines Kunden in verschiedenen angegebenen Zeiträumen.

5. Customer Budget
   Bestimmt die Absatzmenge, den Umsatz, die Kosten, die Marge und den durchschnittlich erzielten Preis für einen Kunden im Jahre 1996.

6. Product Budget
   analog zu Customer Budget, aber für ein Produkt.

7. Forecast Analysis
   Bestimmt die voraussichtlichen Absatzmengen, Umsätze, etc. für einen Zeitraum in Jahre 1996.

8. Budget Performance
   Führt die „Budget vs. Actual-Szenarienanalyse “ und eine Vergleichsanalyse zwischen dem aktuellen und dem vergangenen Jahr durch.

9. Forecast Performance
   Führt die „Forecast vs. Actual-Szenarienanalyse “ und eine Vergleichsanalyse zwischen dem aktuellen und dem letzten Monat durch.

10. Ad Hoc
    Liefert eine der o. g. Maßzahlen.

Query Mix					
Nr.	Abfragename	Anteil	Nr.	Abfragename	Anteil
1	Channel Sales Analysis	10%	6	Product Budget	5%
2	Customer Margin Analysis	10%	7	Forecast Analysis	15%
3	Product Inventory Analysis	15%	8	Budget Performance	20%
4	Time Series Analysis	3%	9	Forecast Performance	15%
5	Customer Budget	5%	10	Ad Hoc	2%

Zur Messung der Leistung eines OLAP-System dient bei APB-1 das Maß AQM („Analytical Queries per Minute“).
AQM ist definiert als der Quotient wie folgt definiert:

$$AQM = \frac{Anzahlderausgeführten Abfragen * 60}{Gesamzeitbedarf}$$

Der Gesamtzeitbedarf beinhaltet neben der reinen Rechenzeit für die Abfrageauswertung auch den Zeitbedarf zum Laden der Daten und den Zeitbedarf evtl. sonstiger Operationen, wie etwa die Datensortierung, da die von APB1GEN erzeugten Daten nicht sortiert sind.
Die Spezifikation von APB-1 schreibt kein spezielles Datenbankschema vor; dies wird damit begründet, das es zum einen verschiedenste Typen von Datenbanksystemen (relational, objektorientiert, etc.) gibt und zum anderen kein allgemein akzeptiertes Datenbank-Design (denormalisiert, 1-NF, 2-NF, usw.) existiert. Somit wäre ein vorgegebenes Datenbankschema von vornherein eine Einschränkung für das zu testende System. Hier lässt man den Benutzern des Benchmarks freie Hand; jedoch muss das Datenbankschema mit den Ergebnissen des Benchmarks veröffentlicht werden.
Wenn APB-1-Benchmarkergebnisse veröffentlicht werden, gehört dazu zwingend ein Testbericht, der die Testumgebung beschreibt. So ist die verwendete Hard- und Software für Server und Client anzugeben, sowie sämtliche vom Benchmarknutzer geschriebenen Quellcodes (z. B. Skripte, etc.). die zur Durchführung des Benchmarks genutzt wurden. Diese Test-Dokumentation muß alle Informationen enthalten, um die Testumgebung und die Benchmarkdurchläufe jederzeit wieder nachbilden zu können.
Ferner muss eine Überprüfung der Ergebnisse durch einen vom OLAP Council autorisierten Auditor erfolgen.

2.2 Transaction Processing Council TPC-D

Der TPC-D Benchmark [TPC98] beschreibt eine Anwendung aus dem Bereich des Decision Support. Dabei werden große Datenmengen (hier mindestens 1 GB) komplexen Analysen unterworfen. Diese Analysen können sehr unterschiedliche inhaltliche Ausprägungen haben, so enthält der TPC-D Benchmark Analysen aus verschiedensten betriebswirtschaftlichen Bereichen:

- Preisgestaltung und Marketing
- Beschaffung und Vertrieb
- Erlösmanagement

- Kundenzufriedenheit
- Marktsegment-Analysen
- Logistik

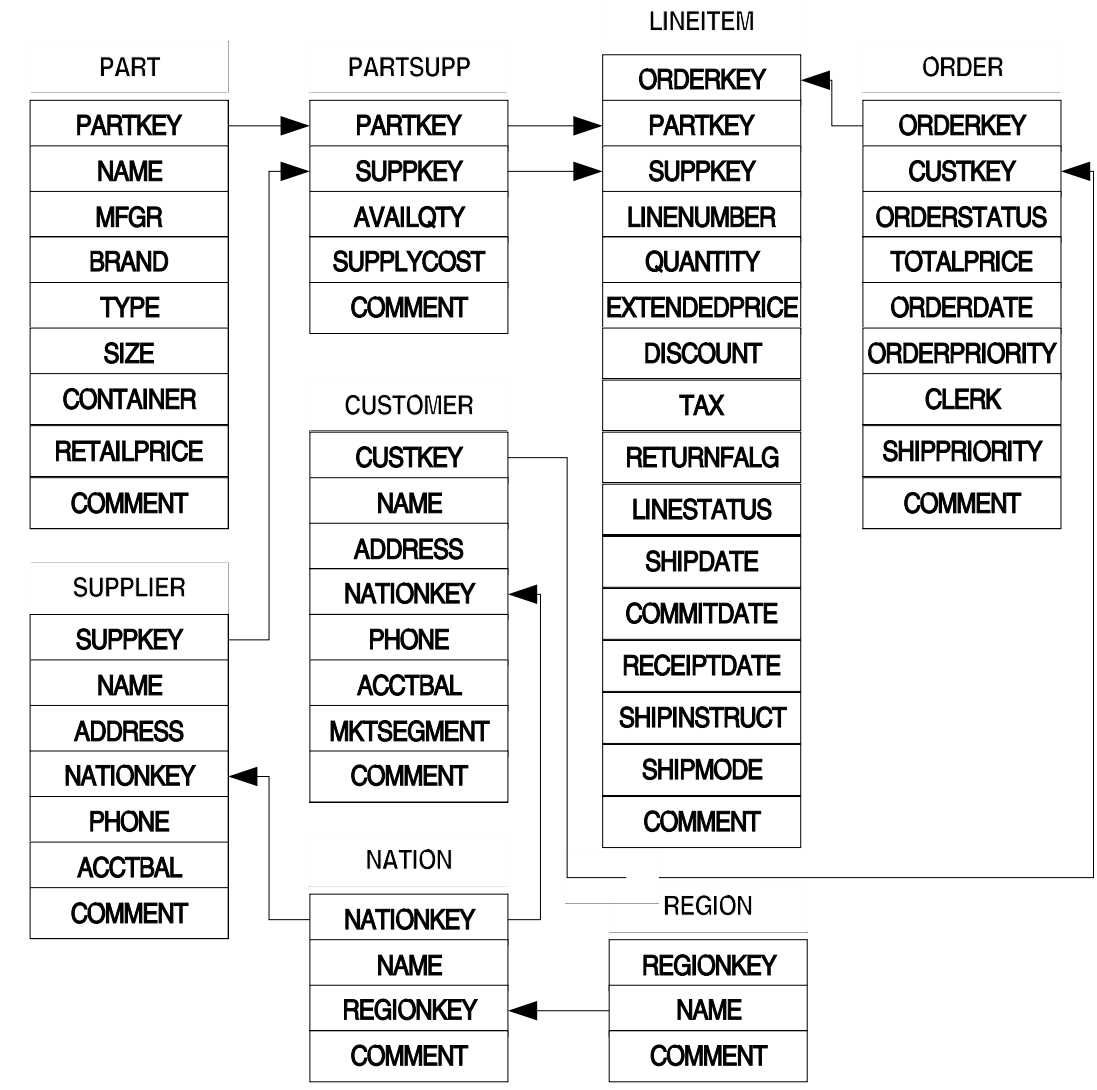

Abbildung 1: Datenbankschema TPC-D

Dem Benchmark liegt das in Abbildung 1 dargestellte Schema zu Grunde. Die nachstehende Tabelle enthält die Tabellen des Datenbankschemas und die Anzahl der Einträge bezogen auf den Skalierungsfaktor (SF, Näheres dazu s. u.).

Tabellengrößen Datenbankschema TPC-D Benchmark		
Tabelle		Anzahl der Tabelleneinträge
Artikel	Part	$200.000 * SF$
Lieferant	Supplier	$10.000 * SF$
Kunde	Customer	$150.000 * SF$
Bestellung	Order	$1.500.000 * SF$
Bestellposition	Lineitem	$6.000.000 * SF$
Zuordnung v. Lief. zu Art.	PartSupp	$800.000 * SF$
Land	Nation	25 (unabh. von SF)
Region	Region	5 (unabh. von SF)

Die Systemumgebung, die dem Benchmark zu Grunde liegt, gliedert sich in zwei Teile, nämlich zum einen dem OLTP-Datenbanksystem, das zur Abwicklung der Geschäftsprozesse genutzt wird, und zum anderen dem DS-Datenbanksystem zur Ausführung des Benchmarks.
Das Modell sieht vor, dass das DS-System in der Lage ist, den Zustand der OLTP-Datenbank (evtl. zeitversetzt) zu übernehmen. Dies kann entweder durch das Kopieren der kompletten Daten („Schnappschuss“) oder durch Nachziehen der Änderungen (regelmäßige Updates) geschehen. Der Standard schreibt nicht vor welche Möglichkeit genutzt werden soll; dies wird damit begründet, dass im Markt beide Lösungen zur Anwendung kommen (siehe dazu [TPC95]).
Das verwendete Datenbanksystem muss zwingend die ACID-Eigenschaften unterstützen, wobei der Datenbankadministrator die Sperrmodi und zusätzliche Scheduling-Parameter für die konkurrente Ausführung von Transaktionen einmalig festlegen darf. Diese Einstellungen müssen in der Testdokumentation festgehalten werden.
Zur Leistungsmessung verwendet TPC-D das Maß „Composite Query-per-Hour Performance“ (QphD@Size), wobei „Size“ für die Größe der Datenbank steht. Diese ergibt sich aus dem gewählten Skalierungsfaktor. Die folgende Tabelle zeigt die erlaubten Skalierungsfaktoren und die resultierende Datenbankgröße (in Gigabyte).

Datenbankgrößen TPC-D Benchmark			
Skalierungsfaktor	Datenbankgröße	Skalierungsfaktor	Datenbankgröße
1	1 GB	300	300 GB
10	10 GB	1.000	1.000 GB
30	30 GB	3.000	3.000 GB
100	100 GB	10.000	10.000 GB

Die Benutzung anderer Datenbankgrößen ist nicht zulässig, da dann die Vergleichbarkeit der Ergebnisse nicht mehr gegeben ist. Von Seiten des TPC wird zusätzlich empfohlen, nur Leistungswerte von Systemen mit gleicher Datenbankgröße zu vergleichen. Aktuelle Benchmarkergebnisse finden sich bei [TPC03a], [TPC03b] und [TPC03c].
Außerdem existiert ein Maß, um die Leistung des Systems in Relation zu den Kosten zu setzen. Es errechnet sich als der Quotient aus den Systemkosten und der gemessenen Leistung (Maß: $\frac{\$}{QphD@Size}$).
Die Kosten umfassen sowohl die Anschaffungskosten für die verwendete Hard- und Software als auch die Wartungskosten für einen 5-Jahres-Zeitraum [Shan98].
Um zu gewährleisten, dass die Datenbank bei jedem Benchmarkdurchlauf die gleiche Datenbasis enthält, wird das Programm DBGEN mitgeliefert, das die notwendigen Datensätze erzeugt. Der Benchmark besteht aus 22 Abfragen, die im Verlauf des Benchmark mit wechselnden Parametern aufgerufen werden, wobei jeder Aufruf eine eigene Transaktion darstellen muss:

1. Pricing Summary Report Query
   Berechnet die gesamten sowie die durchschnittlichen Absatzmengen, rabattierte und nicht rabattierte Umsätze (ohne und mit Steuern) über alle Bestellpositionen in einem Zeitraum von 60 bis 120 Tagen. Die Ergebnisse werden nach Bearbeitungsstatus sortiert und gruppiert.

2. Minimum Cost Supplier Query
   Sucht zu jedem Teil den günstigsten Lieferanten (Name, Adresse etc.) in einer angegebenen Region. Qualifizieren sich mehrere Lieferanten, da sie alle

das Teil zum gleichen Preis anbieten, umfasst das Ergebnis die 100 Lieferanten mit den höchsten Umsätzen. So lässt sich zum einen der günstigste Lieferant finden und zum anderen eine Konzentrierung auf wenige große Kernlieferanten erreichen.

3. Shipping Priority Query
   Liefert die 10 umsatzstärksten Bestellungen, die zum angegebenen Zeitpunkt noch nicht versendet worden sind, in absteigender Sortierung.

4. Order Priority Checking Query
   Liefert die Anzahl der Bestellungen in einem angegebenen Quartal, bei denen mindestens eine Bestellposition nach dem zugesagten Termin geliefert wurde. Zusätzlich erfolgt die Sortierung nach Bestellungspriorität.

5. Local Supplier Volume Query
   Bestimmt für alle Regionen den Umsatz, der erzielt wurde, wenn der Besteller der Artikel und der Teilelieferant im gleichen Land ansässig sind. Dies dient dazu zu entscheiden, ob vor Ort ein Vertriebszentrum eingerichtet werden soll.

6. Forecasting Revenue Change Query
   Analysiert die potenzielle Umsatzsteigerung, die sich durch die Streichung von Kleinstrabatten (1 % und weniger) ergeben würde.

7. Volume Shipping Query
   Berechnet das Transportvolumen zwischen zwei Nationen. Da die Verträge mit Logistikdienstleistern i. d. R. volumenabhängige Rabatte enthalten, können die hier gewonnen Informationen bei der Neuverhandlung von Transportverträgen genutzt werden.

8. National Market Share
   Berechnet die Entwicklung des Marktanteils eines Landes in seiner Region in einem 2-Jahreszeitraum. Die Berechnung erfolgt über Transaktionsvolumina der Lieferanten in der Region.

9. Product Type Profit Measure

Liefert die Umsätze einer Artikelgruppe gegliedert nach Land, Lieferant und Jahr.

10. Return Item Reporting Query
    Berechnet den Geldwert der Rücklieferungen der Kunden in einem Quartal. Es werden die 20 Kunden mit den höchsten Rücklieferungswerten in absteigender Reihenfolge ausgegeben.

11. Important Stock Identification Query
    Liefert den Anteil, den der Lagerbestand eines Lieferanten an der verfügbaren Gesamtmenge eines Teils ausmacht.

12. Shipping Modes and Order Priority Query
    Untersucht die Auswirkungen unterschiedlicher (z. B. kostengünstigerer) Versandwege auf die Termintreue.

13. Consumer Distribution Query
    Bestimmt für jeden Kunden die Anzahl seiner Bestellungen. So lassen sich Hauptabnehmer ermitteln.

14. Promotion Effect Query
    Untersucht die Auswirkungen von Marketing-Kampagnen auf den Umsatz. Es wird der Anteil des Umsatzes berechnet, bei dem die Kaufentscheidung des Kunden in direktem Zusammenhang zur Marketing-Kampagne steht.

15. Top Supplier Query
    Findet den Lieferanten, der den größten Lieferanteil an den im angegebenen Quartal verkauften Artikeln hatte.

16. Parts/Supplier Relationship Query
    Bestimmt, welche Lieferanten das Teil mit den angegebenen Eigenschaften liefern könnten.

17. Small Quantity Revenue Query
    Untersucht, welche Auswirkungen es auf den Umsatz hätte, wenn die Bestellungen von Kleinstmengen zukünftig abgelehnt würden.

18. Large Volume Customer Query
    Liefert die 100 Kunden, deren Bestellmenge die angegebene Grenze überschreitet.

19. Discounted Revenue Query
    Berechnet die in allen Bestellungen insgesamt gewährten Rabatte für die einzelnen Artikel.

20. Potential Part Promotion Query
    Bestimmt die Lieferanten, die große Mengen von Teilen liefern könnten, die für Aktionsartikel nötig sind.

21. Supplier Who Kept Orders Waiting
    Analysiert, welche Lieferanten die zugesagten Liefertermine nicht einhalten konnten, was dann zu einer verspäteten Anlieferung des Produktes beim Kunden führte.

22. Global Sales Opportunity Query
    Untersucht die geographische Verteilung der Bestellungen auf die einzelnen Länder. So lassen sich Kernabsatzmärkte identifizieren.

Diese Abfragen werden während des Benchmarkdurchlaufs in zwei verschiedenen Tests ausgeführt:
Beim Leistungstest „Power Test“ erfolgt die Messung der reinen Ausführungszeit für die Abfragen, da nur ein einzelner Benutzer simuliert wird. Das Leistungsmaß in diesen Tests heißt „TPC-D Power@Size (QppD)“.
Im anschließenden Durchsatztest „Throughput Test“ werden mehrere Benutzer durch parallel ausgeführte „Query Streams“ simuliert. Ihre Anzahl in Abhängigkeit vom Skalierungsfaktor zeigt nachstehende Tabelle.

Skalierungsfaktor	Query Streams	Skalierungsfaktor	Query Streams
1	2	300	6
10	3	1.000	7
30	4	3.000	8
100	5	10.000	9

Beim Durchsatztest erfolgt die Leistungsmessung mittels des Maßes „TPC-D Throughput@Size (QthD)“. Aus diesen beiden Maßen lässt sich dann die Gesamtleistung errechnen. Sie ist wie folgt definiert:

$$QphD@Size = \sqrt{Power@Size * Throughput@Size}$$

Für die Publikation von Benchmarkergebnissen gelten bei TPC-D ähnliche Regelungen wie bei dem APB-1-Benchmark. Zusätzlich sind jedoch die Kosten der verwendeten Soft- und Hardware detailliert (bspw. inklusive gewährter Rabatte) aufzuführen. Neben den reinen Anschaffungskosten sind noch die Wartungskosten für einen Zeitraum von fünf Jahren mit einzurechnen.
Mittlerweile existiert ein Nachfolger des TPC-D: der TPC-H-Benchmark [TPC02a]. Er stellt eine Weiterentwicklung des TPC-D dar. Die grundsätzliche Konzeption, also die Aufteilung in ein OLTP- und ein DS-System, das Datenbankschema und die verwendeten Abfragen wurden übernommen. Zur Leistungsmessung dient nun ein neues Maß: „TPC-H Composite Query-per-Hour (QphH@Size)“.
Basierend auf dem TPC-H existiert noch ein zusätzlicher Benchmark, der TPC-R [TPC02b]. Er führt die gleichen Operationen aus wie der TPC-H, jedoch ist es hier gestattet, vorhandenes Wissen über die Abfragen zu deren Optimierung zu nutzen. Damit kommt der Benchmark der Praxis etwas näher.

3 Standards zur Datenintegration

Datenbanksysteme und die darin enthaltenen Daten sind kritische Komponenten für den Erfolg eines Unternehmens. Neue Anwendungsgebiete wie beispielsweise Datamining- oder OLAP-Anwendungen stellen hohe Anforderungen an die Integration der Daten verschiedener, zumeist heterogener Systeme in eine Applikation. Dabei kommt den Metadaten eine entscheidende Bedeutung zu, da sie wichtige Informationen zur erfolgreichen Datenintegration liefern.
Im folgenden werden exemplarisch drei Standards zur Datenintegration vorgestellt. Zum einen „Microsoft OLE DB“ ein Standard der die Integration verschiedenster Datenquellen ermöglicht und zum anderen die „Metadata Interchange Specification“zur Integration von Metadaten. Abschließend wird mit dem „Com-

mon Warehouse Model“ ein Modell vorgestellt, das geeignet ist, um „Enterprise Application Integration (EAI)“ beschreiben zu können.

3.1 Microsoft OLE DB

Die OLE DB-Schnittstelle stellt eine Menge von Objekten für das Common Object Model (COM) zu Verfügung, die einen transparenten Zugriff auf verschiedenste Datenquellen ermöglichen. So verbirgt die Schnittstelle Details wie etwa den Typ der Datenquelle (ASCII-Datei oder Datenbank) oder den Ort der Speicherung vor dem Benutzer.
Der Anbieter der Daten andererseits kann sehr einfach Zugriffsmechanismen auf die Daten realisieren, indem er die spezifizierten Interfaces implementiert. Diese Interfaces gliedern sich in zwei Klassen. Zum einen die Klasse der Minimalanforderungen; sie beschreibt einen Mindestfunktionsumfang, der bereitgestellt werden muss. Darüber hinaus enthält die zweite Klasse Funktionen, die von vielen Nutzern bezogen auf diesen Datenquellentyp als nützlich erachtet werden, und deshalb implementiert werden sollten.
Es wird eine Testumgebung („Conformance Test“) zur Verfügung gestellt, die untersucht, ob die implementierten Interfaces korrekt sind. Dazu wird das Zusammenspiel mit anderen Komponenten innerhalb einer ADO-Umgebung untersucht.
ADO ist zusammen mit OLE DB und ODBC ein Teil der „Microsoft Data Access Components (MDAC)“. Diese bilden zusammen mit weiteren Schnittstellen ein allgemeines Zugriffsmodell auf Datenquellen, die sogenannte „Universal Data Access-Strategie“, deren Ziel laut [MS00] darin besteht die unterschiedlichsten Datenquellen (Datenbanksysteme, Dateisysteme etc.) über einheitliche Schnittstellen miteinander zu verbinden.
In der Version 2.0 von OLE DB erfolgte die Erweiterung des COM-Modells durch die „OLE DB for OLAP“-Schnittstelle. Laut [MS98] ist es das Ziel dieser Schnittstelle, ausgehend von OLE DB Objekten OLAP Funktionalität, also die Verbindung von mehrdimensionalen Datenquellen und Nutzern, bereitzustellen und dies unabhängig von der Art und dem Ort der Datenspeicherung.
Diese Erweiterungen werden schrittweise in die neuen Versionen von Microsoft-

Programmen wie etwa Access, Excel oder SQL-Server integriert.

3.2 Austausch von Metadaten

Oft findet sich in Unternehmen die folgende Situation: Ähnliche Daten sind in verschiedenen Niederlassungen eines Unternehmens auf unterschiedliche Weise (bezogen auf Datenformate, Datenbankschema etc.) erfasst und gespeichert, so dass zur Auswertung verschiedene Datenbankschemata zusammengeführt werden müssen.
Dieses Szenario beschreibt eine der einfachsten Problemstellungen und dennoch gestaltet sich die Integration von Daten aus verschieden Schemata als kompliziert, weil keine Standards zum Austausch der Metadaten von Datenbanksystemen existieren. Um diese Lücke zu füllen, haben sich verschiedene Anbieter von Datenbanken in der „Metadata Coalition“ zusammengeschlossen mit dem Ziel, ein solches Format zu spezifizieren.
In einigen Unternehmen geht man noch einen Schritt weiter: Da man die Wichtigkeit von Metadaten erkannt hat, versucht man unternehmensweite Metadaten-Management-Systeme zu installieren, um Metadaten zu verwalten, zu ändern und als langfristiges Ziel eine eigene Strategie zur Beschreibung und Nutzung von Metadaten umzusetzen. Da jedoch in keinem Unternehmen eine homogene Systemumgebung – und dies sowohl in Bezug auf Hard- als auch auf Software, existiert – ist es erforderlich, Metadaten unterschiedlichster Systeme austauschen zu können.
Nach [Hamm98] gliedert sich eine Metadaten-Strategie in die folgenden drei Schritte:

1. Definition eines Metadatenmodells
   Die Metadata Coalition beschreibt in ihrer „Metadata Interchange Specification (MDI)“ [Bord] eine Aufteilung des Modells in zwei Komponenten:

   - Application Metamodel
     Dieses Modell beschreibt die Tabellen und sonstigen Datenstrukturen zur Speicherung der Nutzdaten, also der Metadaten, die ausgetauscht werden sollen.

- MetaData Metamodel
  Definiert verschiedene Klassen von Werkzeugen (z. B. für Replikation, Datenextraktion, Abfragen, ...) und deren allgemeine Eigenschaften und dies unabhängig von einem konkreten „Application Metamodel“.

2. Auswahl der Software
   Zusammenstellung der Werkzeuge, die zur Analyse, Verwaltung und Verteilung von Metadaten genutzt werden sollen

3. Definition und Umsetzung von Richtlinien zur Metadatenverwaltung
   Regelt u. a. die Verantwortlichkeiten und Kompetenzen der mit der Metadatenverwaltung betrauten Mitarbeiter. Ferner werden Ansprechpartner (i. d. R. die Administratoren) für die beteiligten Datenbanksysteme bestimmt. Hier können noch weitere unternehmensspezische Punkte (z. B. Dokumentationspflichten etc.) festgelegt werden.

Die Metadata Coalition ging bei der Definition des Austauschformates von folgenden Prämissen aus:

- Bidirektionaler Austausch von Metadaten
  Daraus folgt die Anforderung, dass ein System Metadatenbeschreibungen sowohl importieren als auch exportieren kann. Üblicherweise erfolgt die Datenintegration durch die Spezifikation von Abbildungen („Mapping“) zwischen dem Quell- und dem Zielsystem mittels CASE-Werkzeugen. In diesem Bereich wird eine engere Koppelung zwischen den CASE-Werkzeugen und den Programmen zur Datenextraktion angestrebt. Sollte sich während der Datenextraktion herausstellen, dass das Mapping einen Fehler enthält, so soll der Benutzer diesen Fehler im Extraktionswerkzeug beheben. Dies gibt die Änderung dann an das CASE-Werkzeug weiter.

- Geringer Implementierungsaufwand
  Die Spezifikation soll so gestaltet sein, dass sie mit wenig Entwicklungsaufwand in bestehende Produkte integriert werden kann. Dies war einer der Gründe, warum man sich für ASCII-Dateien als Austauschformat entschlossen hat. Diese bieten darüber hinaus den Vorteil der leichten Portierbarkeit über heterogene Systemgrenzen hinweg.

Auch war es von vornherein nicht beabsichtigt, eine umfassende Spezifikation vorzulegen. Vielmehr ging es darum, in kurzer Zeit eine praktikable Lösung vorzulegen und diese iterativ im Laufe der Zeit zu erweitern. Die Metadata Coalition schloss sich im Jahre 2000 mit der Open Management Group (OMG) zusammen, um die konkurrierenden Standards beider Organisationen in einem Standard zusammenzufassen und damit eine größere Marktakzeptanz erreichen zu können. Der von der Metadata Coalition vorgeschlagene Standard ging in die neue Version des „Common Warehouse Metamodel (CWM)“ der OMG ein.
[Hamm98] beschreibt noch zusätzliche Aspekte, die in den verfügbaren Verwaltungsystemen von Metadaten nur unzureichend unterstützt werden, für Datamining und Datawarehousing jedoch sehr wichtig sind:

- verschiedene Abstraktionsebenen
  Beim Zugriff auf die Daten eines Datenbanksystems besteht die Möglichkeit, durch Aggregation verschieden detailreiche Sichten auf den Datenbestand zu erzeugen. So lassen sich beispielsweise alle Aufträge eines Kunden einzeln oder aber summiert ausgeben. Solche Möglichkeiten zur Abstraktion sollten auch bei Metadaten möglich sein, da etwa der Entwickler einer Analyse-Anwendung eine andere Datensicht benötigt als der Endnutzer.

- Spielregeln
  Leider bieten die heutigen Systeme keine gemeinsame Möglichkeit zur Modellierung von Spielregeln von betriebswirtschaftlichen Anwendungen („Business Rules“). Da jedoch solche Regeln ein wichtiger Bestandteil von Systemen sind, schränkt dieser Umstand die Austauschbarkeit von Metadaten ein.

3.3 OMG Common Warehouse Metamodel

Das Common Warehouse Metamodel hat zum Ziel, einen ganzheitlichen Ansatz zur Modellierung und Verwaltung von Unternehmensdaten zur Verfügung zu stellen. Ausgangspunkt ist die Erkenntnis, dass die im Unternehmen vorhandenen Daten besser genutzt werden müssen, um Erfolgspotenziale und Chancen zu erkennen und dadurch langfristig das Überleben des Unternehmens im Markt und

den finanziellen Erfolg zu sichern.
CWM ist der Versuch, die in den beiden vorherigen Abschnitten vorgestellten Ansätze (Datenintegration und Metadatenintegration) zu verknüpfen. Der Datenaustausch geschieht bei CWM über XML und ist so leicht über Systemgrenzen hinweg realisierbar.
Dies umfasst nach [Iyen00] den kompletten Lebenszyklus von Unternehmensdaten, wobei hier unter Unternehmensdaten sowohl Nutzdaten als auch Metadaten verstanden werden. Der Lebenszyklus von Daten umfaßt u. a. Extraktion, Transformation, Übertragung, Integration und Analyse.
Dem Modell liegt ein 4-Schichten-Modell zu Grunde:

- Foundation Layer
  Diese Schicht bildet die Basis des Systems; eine der Hauptkomponenten ist UML als Modellierungswerkzeug.

- Resource Layer
  Auf dieser Ebene liegen die Datenmodelle der verwendeten Datenquellen. Diese Ebene unterstützt verschiedenste Formen von Datenquellen, seien es Dateien oder verteilte Datenbanksysteme.

- Analysis Layer
  Diese Ebene bietet verschiedene Metamodelle von Diensten auf den Elementen der Ressourcen-Ebene. So erlaubt etwa das „Transformation Metamodel“ die Definition von Abbildungen zwischen Quell- und Zielsystemen. Das OLAP Metamodel erlaubt eine Cube-basierte Datensicht, unabhängig von der physischen Speicherung in einem Datenbanksystem.

- Management Layer
  Dient zur Steuerung („Scheduling“) und Überwachung der vom System bearbeiteten Aufträge.

4 Zusammenfassung

Die hier vorgestellten Technologien aus völlig verschiedenen Bereichen haben gezeigt, wie wichtig herstellerunabhängige Standardisierungen sind.

Im Benchmarkbereich können sie wenigstens ein gewisses Maß an Vergleichbarkeit der Ergebnisse verschiedener Hersteller sicherzustellen. Hier spielt zusätzlich der Aspekt der Auditierung eine Rolle, um ein größtmögliches Maß an Fairness sicherzustellen.
Im Bereich der Datenintegration sind Standards von entscheidender Bedeutung, weil in Unternehmen heterogene Systemumgebungen die Regel und nicht die Ausnahme sind. Mit dem CWM von OMG steht ein Modell zur Verfügung, das es ermöglichen kann, dem nahe zu kommen, was unter dem Schlagwort „Enterprise Application Integration (EAI)“ als Idealbild vernetzter Datenquellen gilt.

Literatur

[Bord] Bordon, Rebecca; The Meta Data Interchange Specification; The Data Adminstration Newsletter;
http://www.tdan.com/

[Burg00] Burgess, Gary; What is the TPC Good For? or, the Top Ten Reasons in Favor of TPC Benchmarks;
http://www.tpc.org/information/other/articles/TopTen.asp

[Hamm98] Hammer, Katherine; Issues in Metadata Exchange; The Journal Of Open Computing
http://www.uniforum.org/journal/MetadataIssues.html

[Iyen00] Iyengar, Sridhar; CWM Audio Briefung: The Key to Integrating Business Intelligence;
http://www.omg.org/news/releases/pr2000/cwm/whitepaper.html June 2000

[MS00] Microsoft Corporation; Choosing Your Data Access Strategie;
http://msdn.mircosoft.com/library/en-us/dnole/html/choosing

[MS98] Microsoft Corporation; OLE DB for OLAP: Frequently Asked Questions;
http://msdn.mircosoft.com/library/en-us/dnole/html/msdn_oledn; July 1998

[Nola99] Nolan Carl; Introduction to Multidimensional Expressions (MDX);
Microsft Corporation; http://www.msdn.microsoft.com/library/en-us/dnolap/html/intromdx.asp; August 1999;

[OLAP98] OLAP Council; APB-1 OLAP Benchmark Release II;
http://www.olapcouncil.org/research/APB1R2_spec.pdf

[OMG00] Object Management Group; Competing Data Warehouse Standards to Merge in the OMG;

http://www.omg.org/news/releases/pr2000/2000-09-25a.html
Press Release 25.09.2000

[Shan98] Shanley, Kim; History and Overview of the TPC http://www.tpc.org/.... Februar 1998

[TPC03a] Complete TPC-C Result List
http://www.tpc.org/tpcr/results/tcpr_results.asp; Juni 2003

[TPC03b] Top Ten TPC-H Performance
http://www.tpc.org/tpch/results/tcph_perf_results.asp; Juni 2003

[TPC03c] Complete TPC-D Result List
http://www.tpc.org/tpcd/results/tcpr_results.asp; Juni 2003

[TPC02a] Transaction Processing Council; TPC Benchmark H Standard Specification Revision 2.0.0;
http://www.tpc.org/tpch/spec/tpch2.0.0.pdf, 2002

[TPC02b] Transaction Processing Council; TPC Benchmark R Standard Specification Revision 2.0.0;
http://www.tpc.org/tpcr/spec/tpcr2.0.0.pdf, 2002

[TPC98] Transaction Processing Council; TPC Benchmark D Standard Specification Revision 2.1;
http://www.tpc.org/tpcd/spec/tpcd_current.pdf; 1998

[TPC95] Transaction Processing Council; TPC D Detailed Description; TPC Quarterly Report ; April 1995
http://www.tpc.org/tpcd/detail.asp

[VaSe] Vassiliadis, Panos; Sellis, Timos; A Survey on Logical Models for OLAP Databases;
Department of Electrical and Computer Engineering, Computer Science Division, Knowledge and Database Systems Laboratory; Athens