Benchmarks und Standards

Vortrag im Rahmen des Seminars

Business Intelligence Teil I:

OLAP und Datawarehousing

Karl-Christian Pammer

18. Juli 2003

Überblick

- OLAP-Benchmarks
  - Motivation
  - TPC-D
  - APB-1
- Standards zur Datenintegration
  - Motivation
  - Integration von operationalen Daten
    - OLE DB
  - Integration von Metadaten
    - MDIS

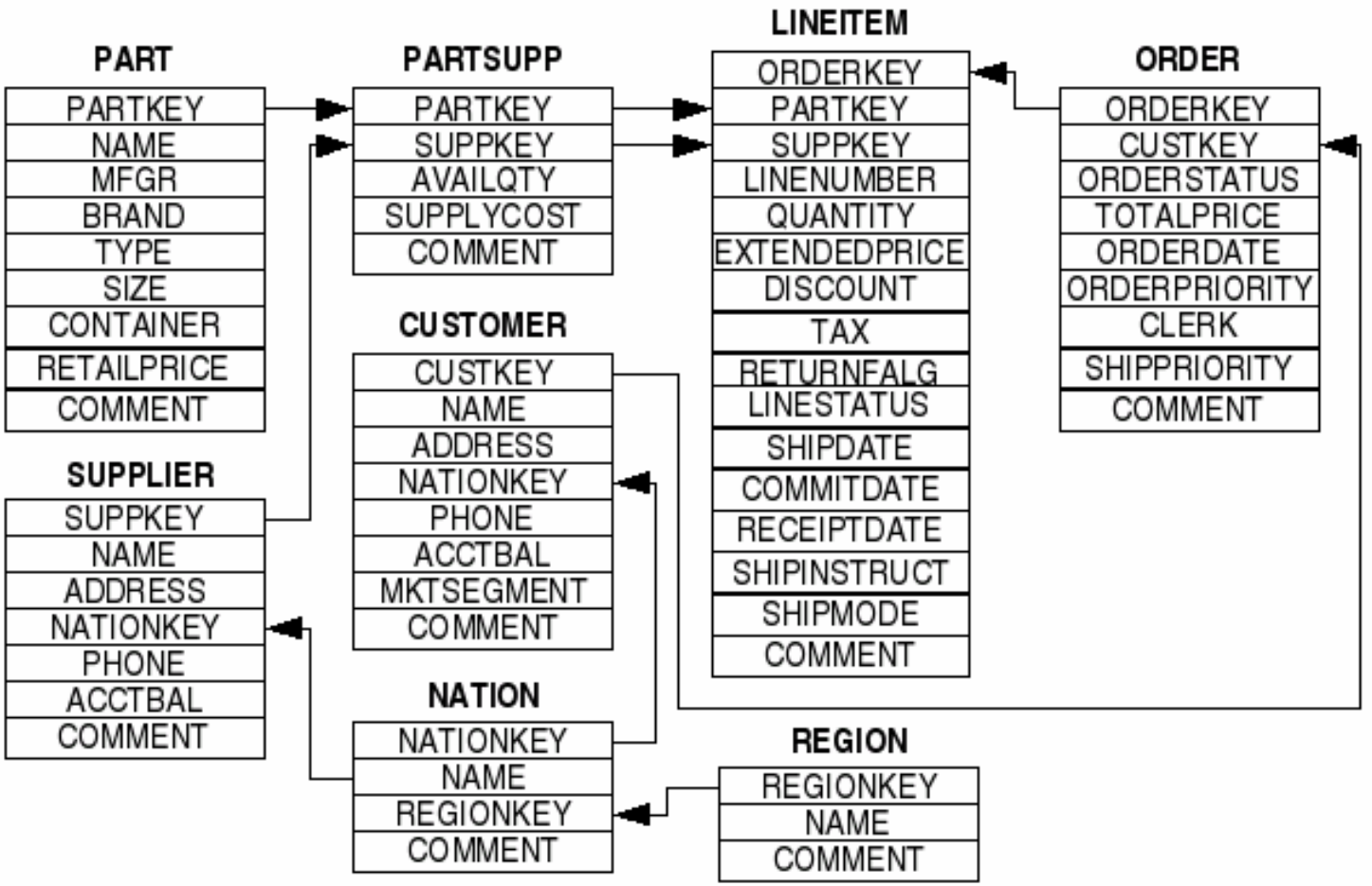
Benchmarking und Benchmarketing

- Hersteller-Benchmarks
  - Unzureichend dokumentiert
  - Wenig repräsentativ
  - Systemorientiert
- Standard-Benchmarks
  - Benchmark-Spezifikation
  - Dokumentationspflichten
  - Auditierung

TPC-D: Allgemeines

- Decision Support Benchmark
- Spezifiziert vom Transaction Processing Council
  - 1995: Version 1.0
  - 1998: Version 2.1 (aktuelle Version)
- Systemmodell
  - Multi-user Datenbanksystem
  - Aufteilung in OLTP- und DS-System
  - ACID-Transaktionen
- Datenbasis
  - Unternehmensdaten
  - Datenerzeugung mittels DBGEN

TPC-D: Datenbankschema



TPC-D: Anwendungsszenario

- Analysen aus 6 Bereichen
  - Preisgestaltung und Marketing
  - Beschaffung
  - Erlösmanagement
  - Kundenzufriedenheit
  - Marktsegmentanalyse
  - Logistik
- Realisierung
  - 22 Analyseabfragen (Q1 bis Q22)
  - 2 Aktualisierungsabfragen (RF1, RF2)
  - Vorgegebene Datenbankgrößen (1GB bis 3.000GB)

TPC-D: Abfragen

- Q17: „Small-Quantity-Order Revenue Query“
  - Analyse
    - Welche Erlösminderung ergibt sich, wenn keine Bestellungen von Kleinstmengen mehr akzeptiert werden?
  - Realisierung

```
SELECT   sum(l_extendedprice) / 7.0 as avg_yearly
FROM     lineitem, Part
WHERE    p_partkey = l_partkey
         AND …
         AND l_quantity < (
            SELECT   0.2 * avg(l_quantity)
            FROM     lineitem
            WHERE    l_partkey = p_partkey
         );
```

TPC-D: Durchlauf und Leistungsmaße

- Durchlauf
  - „Power Test"
    - Einen Query-Stream
    - Einen Refresh-Stream
    - Maß
      - TPC-D Power@Size (QppD)
  - „Throughput Test"
    - Mehrere Query-Streams
    - Maß
      - TPC-D Throughput@Size (QthD)
- Gesamtmaße:
  - „TPC-D Composite Query-per-Hour (QphD)"
  - „TPC-D Price-per-QphD@Size"

TPC-D: Dokumentation

- Durchlauf
  - Datenbankgröße
  - Ausführungszeiten (RF1, RF2, Q1 bis Q22)
    - Power Test
    - Throughput Test
- Testumgebung
  - Verwendete Hardware
  - Verwendete Software
  - Systemkosten (inkl. 5 Jahre Wartung)
- Ergebnis der Auditierung

TPC-D: Nachfolger

- TPC-H
  - 1999: Version 1.0
  - 2002: Version 2.0.0 (aktuelle Version)
  - Nachfolger des TPC-D
  - Geänderte Leistungsmaße
- TPC-R
  - Basiert auf TPC-H
  - Optimierung der Anfragen

TPC-D: Ergebnisse

Com-Pany	System	TPC-Power QppD	TPC-Through-put QthD	Composite Query-Per-Hour QphD	Price Per QphD [US-$]	System Availability	Database	Date Submitted
Teradata	WorldMark 5200	133,966	13,756	42,928	440	08/10/99	NCR Teradata V2R3.0	02/15/99
HP	NetServer LXr 8000	8,124	1,324	3,280	162	05/18/99	Oracle8i 8.1.5.1.1	02/11/99
Sequent	NUMA-Q 2000	1,854	572	1,030	3,999	02/27/98	Oracle8 8.0.4	01/20/98

- Datenbankgröße: 300 GB

APB-1: Allgemeines

- OLAP Benchmark
- Spezifiziert vom OLAP Council
  - 1996: Release I
  - 1998: Release II (aktuelle Version)
- Systemmodell
  - Client-Server-Modell
  - Datenhaltung und Berechnungen erfolgen serverseitig
- Datenbasis
  - Vertriebs- und Marketingdaten
  - Datenerzeugung mittels APB.EXE

APB-1: Datenbankschema

- Kein vorgegebenes Schema
- APB.EXE erzeugt ASCII-Dateien
- 5 hierarchische Objekttypen
  - Product
    - Top
    - Division
    - Line
    - Family
    - Group
    - Class
    - Code
  - Customer
    - Top
    - Retailer
    - Store
  - Channel
    - Top
    - Base
  - Scenario
    - Budget
    - Actual
    - Forecast
  - Time
    - Inventory
    - Aggregations
      - Quarterly
      - Yearly

APB-1: Anwendungsszenario

- Analysen aus 7 Bereichen
  - Absatzkanal-Analyse (10%)
  - Margen-Analyse (10%)
  - Bestand-Analyse (15%)
  - Zeitreihen-Analyse (3%)
  - Budget-Analyse (30%)
  - Vorhersage-Analyse (30%)
  - Ad Hoc Anfrage (2%)
- Realisierung
  - 10 Abfragen
  - Nur Lesezugriffe

APB-1: Abfragen

- Q1: „Channel Sales Analysis“
  - Analyse
    - Absatzmengen und Umsätze für einen Absatzkanal
    - Parameter
      - Artikel (?product)
      - Kunde (?customer)
      - Absatzkanal (?channel)
      - Zeitraum (?time)
  - Formale Beschreibung

    get UNITS SOLD, DOLLAR SALES, AVERAGE PRICE
    by SCENARIO = „ACTUAL“
    by PRODUCT = <children(?product)>
    by CUSTOMER = <children(?customer)>
    by CHANNEL = <?channel>
    by TIME = <children(?time)>

APB-1: Durchlauf und Leistungsmaße

- 6 Schritte
  1. Erzeugen der Grunddaten
  2. Initialisierung der Datenbank
  3. Erzeugen der inkrementellen Daten
  4. Laden der inkrementellen Daten / Vorberechnungen
  5. Erzeugen der Abfragen
  6. Ausführung der Abfragen
     - Anzahl der Abfragen pro Query-Stream abhängig von der Größe der Channel-Tabelle
- Leistungsmaß
  - „Analytical Queries per Minute“
  - Berechnet aus den Zeiten der Schritte 4 bis 6

APB-1: Dokumentation

- Ähnlich zu TPC-D
- Zusätzlich gefordert
  - Datenbankschema
  - Programmcode / Skripte, die genutzt wurden für:
    - Die Erzeugung der Datenbank
    - Das Laden der Daten in die Datenbank
    - Eventuelle Vorberechnungen
    - Die Ausführung der Abfragen
    - Die Client-Seite
  - Anzahl der simulierten Benutzer

APB-1: Ergebnisse

	Company	System	AQM	Database	Operating System	Date
1	HP	4x HP rp7400 Server	85,719	Oracle 9.2.0.2.0	HP-UX 11i	12/09/02
2	Sun	Sun Enterprise 450 Server	8,073	Oracle Express 6.1	Sun Solaris	05/28/98

Benchmarks: Zusammenfassung

- Versuch Systeme vergleichbar zu machen
- Benchmarks simulieren „Best Practices“
- Kein Ersatz für anwendungsbezogene Evaluation

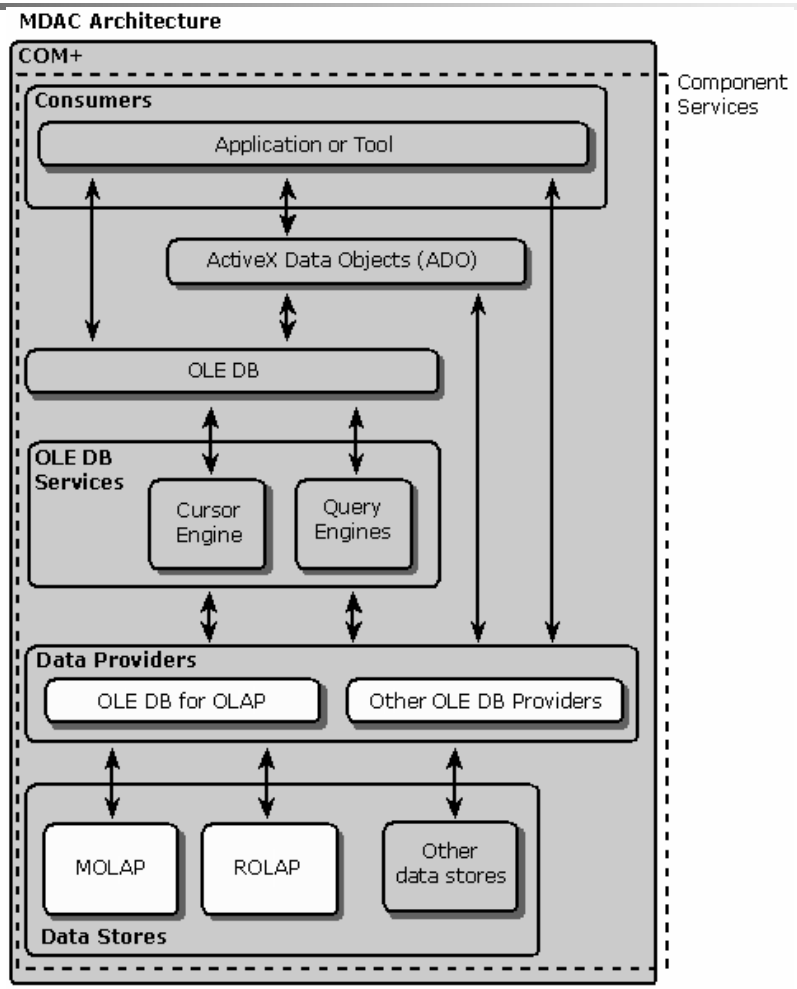
Standards: Motivation

- Klassifizierung der Datenintegration
  - Integration von operationalen Daten
  - Integration von Metadaten
- Probleme der Datenintegration
  - Heterogene Hard- und Software-Systeme
  - Unterschiedliche Erfassung ähnlicher Daten bezogen auf
    - Datentypen
    - Datenformate
    - Datenbankschema

Standards: Microsoft OLE DB

- Erweiterung des „Common Object Model (COM)“
- Teil der „Microsoft Data Access Components (MDAC)“
- „Middleware“ zwischen Anwendern und Datenquellen
- Ziele
  - Verknüpfung verschiedenster Arten von Datenquellen
  - Ortstransparenz
  - Standardisierte Zugriffsschnittstellen
    - Mindestfunktionalität
    - Erwünschte Funktionalität
- Version 2.0: Erweiterung um OLAP-Funktionalität („OLE DB for OLAP“)

Standards: MDAC Architektur



Standards: Integration von Metadaten

- Implementierung einer Metadaten-Management-Strategie
  - Definition eines Metadaten-Modells
    - z. B. „Metadata Interchange Specification (MDIS)“ der Metadata Coalition
  - Software-Auswahl
    - Werkzeuge zur Verwaltung, Verteilung etc.
  - Definition und Umsetzung von Richtlinien
    - Kompetenzen und Verantwortlichkeiten
    - Ansprechpartner
    - Dokumentationspflichten
    - Weitere unternehmensinterne Aspekte

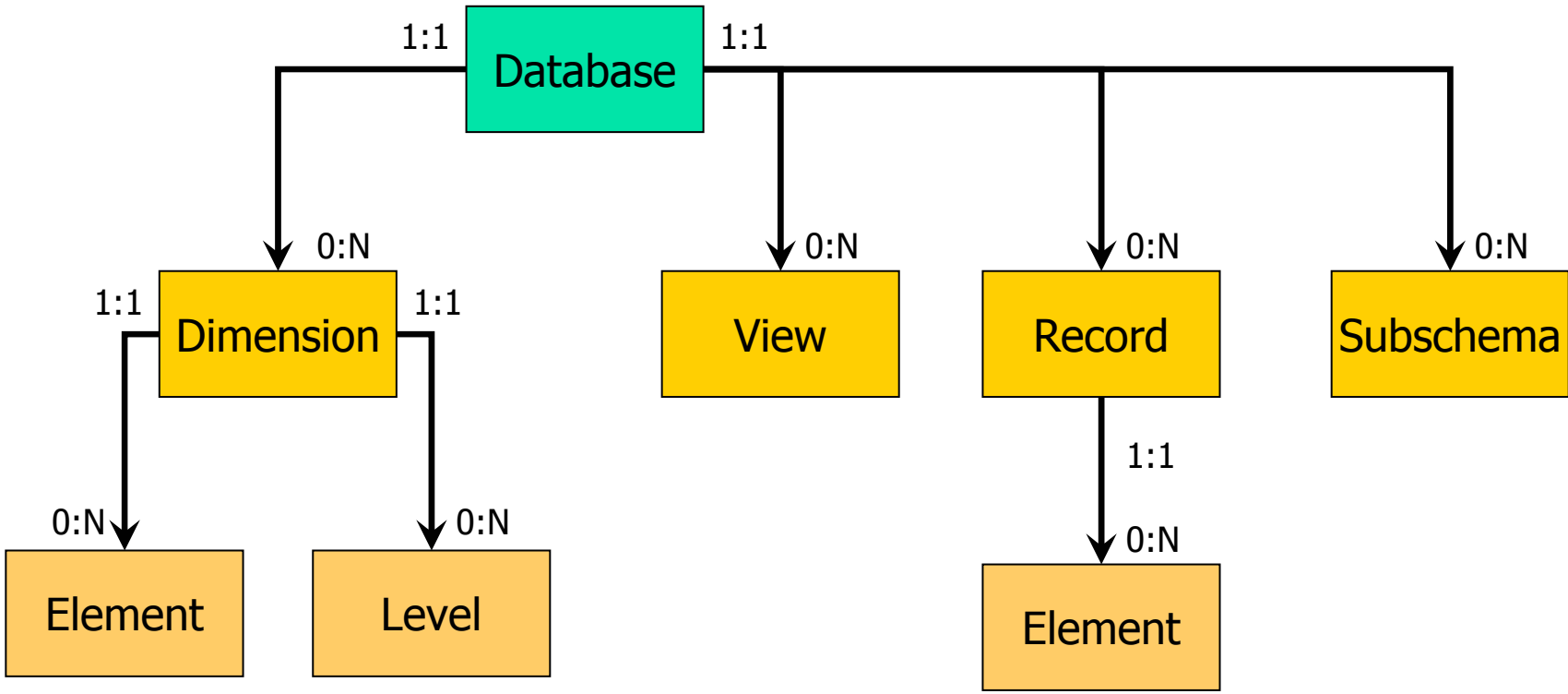
Standards: Metadata Interchange Specification

- Herausgegeben von der „Metadata Coalition“
  - 1995: Version 1.0
  - 1997: Version 1.1
  - Mittlerweile Teil des „Common Warehouse Models (CWM) der „Object Management Group (OMG)“
  - ASCII-Dateien-basiertes Austauschformat
- Prämissen
  - Keine allumfassender Standard angestrebt
  - Einfach zu implementieren

Standards: Metadata Interchange Specification

- Metadaten-Modell
  - Application Metamodel
    - Datenstruktur zur Speicherung der Metadaten
  - Metadata Metamodel
    - Modellierung der Gemeinsamkeiten von Werkzeugklassen (discovery tools, extraction tools, replication tools, …)
    - Fehlermodell
    - Unabhängig vom Application Metamodell

Standards: MetaObjects des MDIS Application Metamodel



Standards: MetaObjects des MDIS Application Metamodell

- Database
  - Alle Arten von Datenquellen
    - Relationale-DB
    - Netzwerk-DB
    - Objekt-DB
    - Hierachische-DB
    - Dateien
- Subschema
  - Logische Gruppe von
    - Tabellen
    - Records
    - Objekten
    - Segmenten
    - Dateien
- Record
  - Physische Gruppierung von Elementen
    - Tabelle
    - Segmente
    - Objekte
- Element
  - Beschreibung der physischen Repräsentation
    - Spalten einer Tabelle
    - Attribute und Klassen-methoden
- Dimension
  - Menge von Daten
  - Zugriff über „HyperCube-Koordinaten“

Standards: MDIS Beispiel

- Datenbank für Abteilungen und Mitarbeiter
  - ABT(A_ID, A_NAME, A_CHEF)
  - MIT(M_ID, M_NAME, …)

BEGIN DATABASE

Identifier „001“

ServerName „Zentralserver1“

OwnerName „DB-Admins“

DatabaseName „Firma“

DatabaseExtendedType „IBM DB2 6.1“

DatabaseType „Relational“

Standards: MDIS Beispiel

BEGIN RECORD
- Identifier „002“
- RecordName „Abt“
- RecordType „Table“
- BEGIN ELEMENT
  - Identifier „003“
  - ElementName „A_ID“
  - ElementDataType „UNSIGNED-INTEGER“
  - ElementKeyPosition „1“
  - ElementNulls „F“
  - ElementOrdinality „1“
- END ELEMENT
- ...

END RECORD

weitere Identifier:
- A_NAME: „004“
- A_CHEF: „005“
- Tabelle MIT: „006“
- M_ID: „007“

BEGIN RELATIONSHIP
- Identifier „008“
- SourceObjectIndentifier „005
- TargetObjectIdentifier „007“
- RelationShipType „EQUIVALENT“
- RelationOrdinality „1:1“

END RELATIONSHIP

END DATABASE

Standards: Zusammenfassung

- Datenintegration ist zentrale Aufgabe
- Integration von Metadaten
  - Bis jetzt nur erste Ansätze
  - Offene Punkte u. a.
    - Definition verschiedener Abstraktionsebenen
    - Herstellerunabhängige Modellierung von Spielregeln
- Integration von operationale Daten
  - Realisierung über Middleware-Mechanismen
  - Stärkere Kopplung mit Metadatenintegration wünschenswert

Vielen Dank für die Aufmerksamkeit

Maßberechnung

- TPC-D
  - Power Test

$$Power@Size = \frac{3600 * SF}{\sqrt[24]{\prod_{i=1}^{22} T(Q_i) * \prod_{j=1}^{2} T(RF_j)}}$$

  - Throughput Test

$$Throughput@Size = \frac{|QueryStreams| * 22 * 3600}{T(Q_{1,..,22}, RF_{1,2})}$$

  - Composite Query-per-Hour

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

- APB-1

$$AQM = \frac{|Queries| * 60}{T(Load) + T(Sort, Calc) + T(Queries)}$$

TPC-H: Ergebnisse nach Leistung für 300 GB Datenbankgröße

	Com-pany	System	QphH	Price per QphH	System Availability	Database	Operating System	Date Submitted
1	HP	Compaq Proliant DL760 x900-64P	12,995	203 US-$	06/20/02	IBM DB2 UDB 7.2	Microsoft Windows 2000 Advanced Server	04/09/02
2	HP	HP AlphaServer ES45 Model 68/100	5,976	401 US-$	06/01/02	Oracle 9i R2 Enterprise Edition	Compaq Tru64 Unix V5.1A/IPK	11/18/02
3	Unisys	Unisys ES7000 Orion 130 Enterprise Server	4,774	208 US-$	03/31/03	Microsoft SQL-Server 2000 Enterprise Edition 64-bit	Microsoft Windows .NET Datacenter Server	10/29/02
4	HP	HP Proliant DL760G2 8P	3,334	71 US-$	05/28/03	Microsoft SQL-Server 2000 Enterprise Edition	Microsoft Windows Server 2003 Enterprise Server	05/28/03
5	Sun	Sunfire V240	1,026	49 US-$	06/23/03	Sybase Sybase IQ 12.5	Sun Solaris 9	06/23/03

TPC-H: Ergebnisse nach Kosten für 300 GB Datenbankgröße

	Com-pany	System	QphH	Price per QphH	System Availability	Database	Operating System	Date Submitted
1	Sun	Sunfire V240	1,026	49 US-$	06/23/03	Sybase Sybase IQ 12.5	Sun Solaris 9	06/23/03
2	HP	HP Proliant DL760G2 8P	3,334	71 US-$	05/28/03	Microsoft SQL-Server 2000 Enterprise Edition	Microsoft Windows Server 2003 Enterprise Server	05/28/03
3	HP	Compaq Proliant DL760 x900-64P	12,995	203 US-$	06/20/02	IBM DB2 UDB 7.2	Microsoft Windows 2000 Advanced Server	04/09/02
4	Unisys	Unisys ES7000 Orion 130 Enterprise Server	4,774	208 US-$	03/31/03	Microsoft SQL-Server 2000 Enterprise Edition 64-bit	Microsoft Windows .NET Datacenter Server	10/29/02
5	HP	HP AlphaServer ES45 Model 68/100	5,976	401 US-$	06/01/02	Oracle 9i R2 Enterprise Edition	Compaq Tru64 Unix V5.1A/IPK	11/18/02