# DATA WAREHOUSING: an introduction

Prof. Sham Navathe
Georgia Institute of Technology
www.cc.gatech.edu/~sham

## Acknowledgement

- Umeshwar Dayal, H.P. labs
- Surajit Chaudhari, Microsoft Research
- Prof. A. Datta and Helen Thomas, Dupree School of Management
- Bill Thomas, CiTi Inc., Dept. of State, Washington D.C.

# Outline

- Introduction to Decision support and OLAP applications
- Technology Evolution
- Terminology: OLAP vs. OLTP
- Data Warehousing Architecture
- General Discussion of DW design and operation
- Multidimensional Modeling
- OLAP processing and tools
- Pointers to products and literature




# Decision Support and OLAP Applications

- Decision Support: Information technology to help the knowledge worker (adjudicator, consulate officer, state department official) make faster and better decisions
  - *How many visas of type B-1 were issued in Asia country by country and for the three most busy months in the year 2001?*
  - *How were visa numbers affected by implementing a new set of restrictions since October 2001?*
- OLAP: On-Line Analytical Processing consists of use of tools for ad-hoc analysis against large data warehouses

# Evolution – Where are we?

- 60's: Batch reports
  - file based
  - hard to find and analyze, repetitive, inconsistent
  - inflexible and expensive, reprogram every new request
- 70's: Heavy use of legacy data model database systems (Hierarchical, e.g., IMS, and Network, e.g., IDMS) for tracking of business transactions



# Evolution – Where are we?

- 80's: PC arrives - Desktop data access and analysis tools
  - query tools, spreadsheets, GUIs
  - easier to use, but only access operational databases
- 90's: Data warehousing with integrated OLAP engines and tools
- 2000's: Personalization engines and e-commerce apps. to work with DW and OLAP/mining tools

# OLTP vs. OLAP

- OLTP : Online Transaction Processing – has been in use to process and record transactions that create new data and update existing information in databases.
- OLAP : Online Analytical Processing

- data is aggregated, warehoused, and then analyzed; users query and generate reports without modifying any data.




# OLTP vs. OLAP

| | OLTP | OLAP |
|---|---|---|
| User | Clerk, IT Professional | Knowledge worker |
| Function | Day to day operations | Decision support |
| DB Design | Application-oriented (E-R based) | Subject-oriented (Star, snowflake) |
| Data | Current, Isolated | Historical, Consolidated |
| View | Detailed, Flat relational | Summarized, Multidimensional |
| Usage | Structured, Repetitive | Ad hoc |
| Unit of work | Short, Simple transaction | Complex query |
| Access | Read/write | Read Mostly |
| Operations | Index/hash on prim. key | Lots of scans |
| # Records accessed | Tens | Millions |
| #Users | Thousands | Hundreds |
| Db size | 100 MB-GB | 100GB-TB |
| Metric | Transaction throughput | Query throughput, response |

# Data Warehouse - definition

- A decision-support database that is maintained separately from the organization's operational (transactional) databases
- A data warehouse is a
  - subject-oriented,
  - integrated,
  - time-varying,
  - non-volatile (static)

  collection of data that is used primarily in organizational decision making



# Data Warehousing Market

Evolved from the database servers and systems market

- Hardware: servers, storage, clients
- Warehouse DBMs
- Tools
- Market growth:
  - $2B in 1995 to $8 B in 1998 (Meta Group)
  - 1.5B in 1995 to $6.9B in 1999 (Gartner Group)
- Systems integration & Consulting
- Already deployed in many industries: manufacturing, retail, financial, insurance, transportation, telecom., utilities, healthcare.

# Why A Separate Data Warehouse?

Performance

- Operational Databases are designed & tuned for known transactions & workloads
- Complex OLAP queries would require full scans of data and degrade performance beyond acceptable level
- Special data organization, access & implementation methods are needed for multidimensional views & queries that conventional database management systems (DBMSs) do not provide

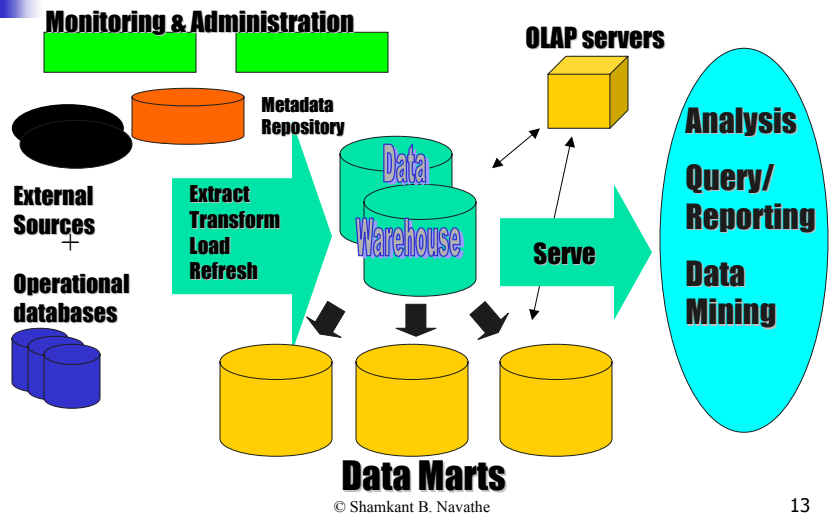



# Why A Separate Data Warehouse?

Function

- Historical data: Decision support requires historical data, which operational Databases do not typically maintain
- Data consolidation: Decision support requires consolidation (aggregation, summarization) of data from many heterogeneous sources: operational databases, external sources
- Data quality: Different sources typically use inconsistent data representations, codes, and formats which have to be reconciled

# Data Warehousing Architecture







# Three-Tier Architecture

- Warehouse database server
  - Almost always a relational DBMS; rarely flat files
- OLAP servers
  - extended relational DBMS that maps operations on multidimensional data to standard relational operations.
- Clients

  House a variety of tools for the decision worker
  - Query and reporting tools.
  - Analysis tools
  - Data mining tools (e.g., trend analysis, prediction)

# Data Warehouse vs. Data Marts

- Enterprise Warehouse: collects all information about subjects (cases, adjudicators, visas, posts, countries) that span the entire application
  - Requires extensive business modeling
  - Different countries can use similar models
- Data Marts: Departmental subsets that focus on selected subjects
  - Adjudication data mart that focuses on cases for adjudication
  - Faster roll out, less comprehensive, less expensive
- Virtual Warehouses: Summary Views over Operational Data
  - Easier to build – as snapshots
  - Used over operational DBs , e.g., banks for read-only




# Traditional DW Design & Operational Process

- Define architecture. Do capacity planning
- Integrate Database and OLAP servers with client tools
- Design warehouse schema, views
- Design physical warehouse organization: data placement, partitioning, access methods
- Connect sources: gateways, ODBC drivers, wrappers
- Design & implement scripts for data extract, load refresh
- Define metadata and populate repository
- Design & implement end-user applications
- Roll out warehouse and applications
- Monitor the warehouse

# OLAP for Decision Support

- Goal of OLAP is to support ad-hoc querying for the "business" analyst
- Business analysts are familiar with spreadsheets
- OLAP extends spreadsheet-like analysis to work with warehouse data
  - Works with large data set and creates a summary view
  - Semantically enriched to understand business and application terms (e.g., time, visa status)
  - Combined with reporting features
- *Multidimensional* view of data is the foundation of OLAP



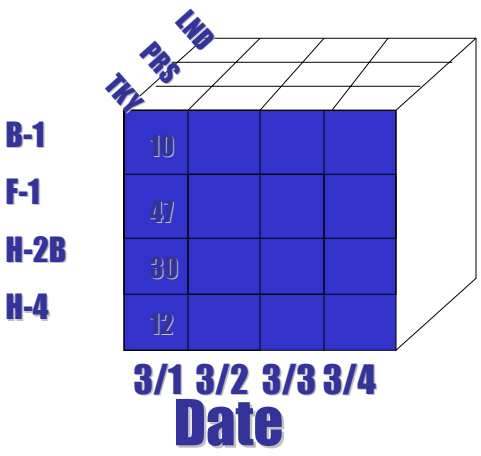

# Multidimensional Data Model

- Database is a set of *facts* (points) in a multidimensional space
- A fact has a *measure* dimension
  - quantity that is analyzed, e.g., number of visas
- A set of *dimensions* on which data is analyzed
  - e.g. , country, consulate, date of issue for a visa
- Each dimension has a set of *attributes*
  - e.g., "Visa" dimension has visa date, visa type, visa category

Attributes of a dimension may be related by partial order

- *Hierarchy*: e.g., post>county>region
- *Lattice*: e.g., date>month>year, date>week>year

# Multidimensional Data



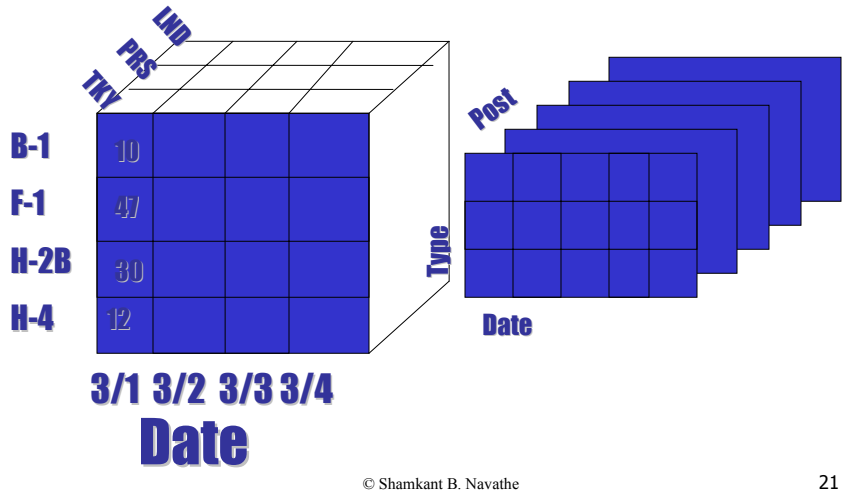

**Visas issued as a function of time, Post and type**




# Operations in Multidimensional Data Model

- Aggregation (*roll-up*)
  - dimension reduction: e.g., total visas by post
  - summarization over aggregate hierarchy: e.g., total post by country , total visas by country and by month
- Selection (*slice*) defines a subcube
  - e.g., no. of visas where post = Istambul and month = 1/2002
- Navigation to detailed data (*drill-down*)
  - e.g., no.of visas issued by post for each month, for top 20% of posts by average visas per post per month
- Visualization Operations (e.g., Pivot)

# A Visual Operation: Pivot (Rotate)







# Approaches to OLAP Servers

- Relational OLAP (ROLAP)
  - Relational and Specialized Relational DBMS to store and manage warehouse data
  - OLAP middleware to support missing pieces
    - Optimize for each DBMS backend
    - Aggregation Navigation Logic
    - Additional tools and services
  - Example: Microstrategy, MetaCube (Informix), Sybase IQ, H-P Intelligent Warehouse
- Multidimensional OLAP (MOLAP)
  - Array-based storage structures
  - Direct access to array data structures
  - Example: Essbase (Arbor), Accumate (Kenan)
- Domain-specific enrichment

# Relational DBMS as Warehouse Server

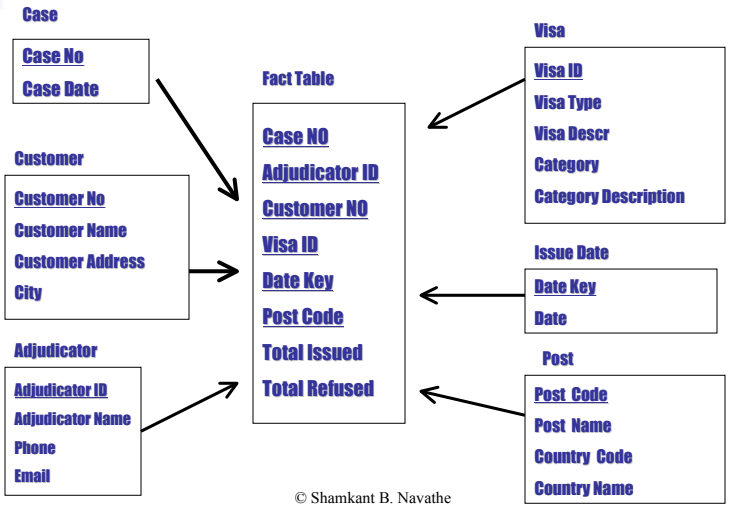
- Schema design is modified for the warehouse
- Specialized scan, indexing and join techniques
- Handling of aggregate views (querying and materialization)
- Supporting query language extensions beyond SQL
- Complex query processing and optimization
- Data partitioning and parallelism




# Warehouse Database Schema

- ER (entity-relationship modeling) based design techniques not appropriate
- Design should reflect multidimensional view
  - Star Schema
  - Snowflake Schema
  - Fact Constellation Schema

# Example of a Star Schema



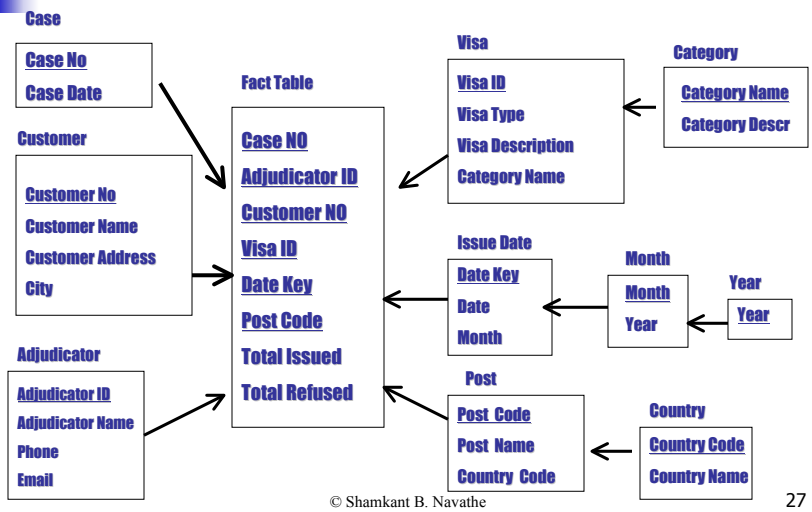




# Star Schema

- A single fact table and a single table for each dimension
- Every fact points to one tuple (row)in each of the dimensions and has additional attributes
- Does not capture hierarchies directly
- Generated keys are used for performance and maintenance reasons
- Fact constellation: Multiple Fact tables that share many dimension tables
  - Example: Projected expense and the actual expense may share dimensional tables

# Example of a Snowflake Schema







# Snowflake Schema

- Represents dimensional hierarchy directly by normalizing the dimension tables
- Easy to maintain
- Saves storage, but is alleged that it reduces effectiveness of browsing (Kimball)

# Populating and Refreshing the Warehouse

- Data extraction from operational systems
- Data cleansing
  - Check for valid adjudicator IDs
- Data transformation
  - Convert from codes to meaningful information
  - Type = Business instead of type = B-1
- Load
  - Sort, summarize, consolidate, compute views, check integrity, build indexes, partition
- Refresh
  - Propagate updates from sources to the warehouse




# Data Extraction

- Direct extract from operational systems
  - Custom for every post or country
  - Efficient once established – One-step DW refresh
  - Changes to operational systems will require changes to DW refresh routines
- Flat-file extract from operational systems
  - Data warehouse design is more independent from post or country-specific applications
  - Lower cost DW
  - Two-step DW refresh

# Data Cleansing

- Why?
  - Data warehouse contains data that is analyzed for business decisions
  - More data and multiple sources could mean more errors in the data and harder to trace such errors
  - Results in incorrect analysis
- Detecting data anomalies and rectifying them early has huge payoffs
  - Match old customer ID numbers with new customer ID #s
  - Match data of visas with data about entries at airports
- Long Term Solution
  - Change business practices and data entry tools




# Data Cleansing Techniques

- Transformation Rules
  - Example: translate "gender" to "sex"
  - Products: Warehouse Manager (Prism), Extract (ETI), Passport (Carleton), ETL flow tools (Sagent)
- Uses domain-specific knowledge to do scrubbing
- Parsing and fuzzy matching
  - Multiple data sources (can designate a preferred source)
  - Products: Integrity (Vality – now Ascential), Trillium Software System 6
- Discover facts that flag unusual patterns (auditing)
  - Some teacher has never received a single complaint
  - Products: WizRule (Hallogram Publishing), QDB, SBStar

# Load

Issues:

- Large volumes of data to be loaded
- Small time window (usually at night) when the warehouse can be taken off-line
- When to build indexes and summary tables
- Allow system administrator to monitor status, cancel suspend, resume load, or change load rate
- Restart after failure with no loss of data integrity



# Refresh

- Issues:
  - When to refresh
    - on every update: too expensive, only necessary if OLAP queries need current data *(e.g., up-the-minute stock quotes)*
    - periodically (e.g., every 24 hours, every week) or after "significant" events
    - refresh policy set by administrator based on user needs and traffic
    - possibly different policies for different sources
  - How to refresh

# Refresh Techniques

- Full extract from base tables
  - read entire source table or database: expensive
  - may be the only choice for legacy databases or files.
- Incremental techniques (related to work on active databases)
  - detect & propagate changes on base tables: replication servers (e.g., Sybase, Oracle, IBM Data Propagator)
    - snapshots & triggers (Oracle)
    - transaction shipping (Sybase)
  - logical correctness
    - computing changes to star tables
    - computing changes to derived and summary tables
    - optimization: only significant changes
  - transactional correctness: incremental load




# Metadata Repository

**METADATA : Data about data**

- Administrative metadata
  - source databases and their contents
  - gateway descriptions
  - warehouse schema, view & derived data definitions
  - dimensions, hierarchies
  - pre-defined queries and reports
  - data mart locations and contents
  - data partitions
  - data extraction, cleansing, transformation rules, defaults
  - data refresh and purging rules
  - user profiles, user groups
  - security: user authorization, access control

# Metadata Repository

- State Department metadata
  - Consular/civil service terms and definitions
  - Sources and ownership of data
  - Policies and accountability
- Operational metadata
  - data lineage: history of migrated data and sequence of transformations applied
  - currency of data: active, archived, purged
  - monitoring information: warehouse usage statistics, error reports, audit trails.




# Warehouse Design Tools

- Creating and managing a warehouse is hard
- Development tools
  - defining & editing metadata repository contents (schemas, scripts, rules).
  - Queries and reports
  - Shipping metadata to and from RDBMS catalog (e.g., Prism Warehouse Manager).
- Planning & analysis tools
  - impact of schema changes
  - capacity planning
  - refresh performance: changing refresh rates or time windows

# Warehouse Management Tools

- Monitoring and reporting tools (e.g., HP Intelligent Warehouse Advisor)
  - which partitions, summary tables, columns are used
  - query execution times
  - for summary tables, types & frequencies of roll downs
  - warehouse usage over time (detect peak periods)
- Systems and network management tools (e.g., HP OpenView, IBM NetView, Tivoli): traffic, utilization
- Exception reporting/alerting tools (e.g., DB2 Event Alerters, Information Advantage InfoAgents & InfoAlert)
- Analysis/Visualization tools: OLAP on metadata



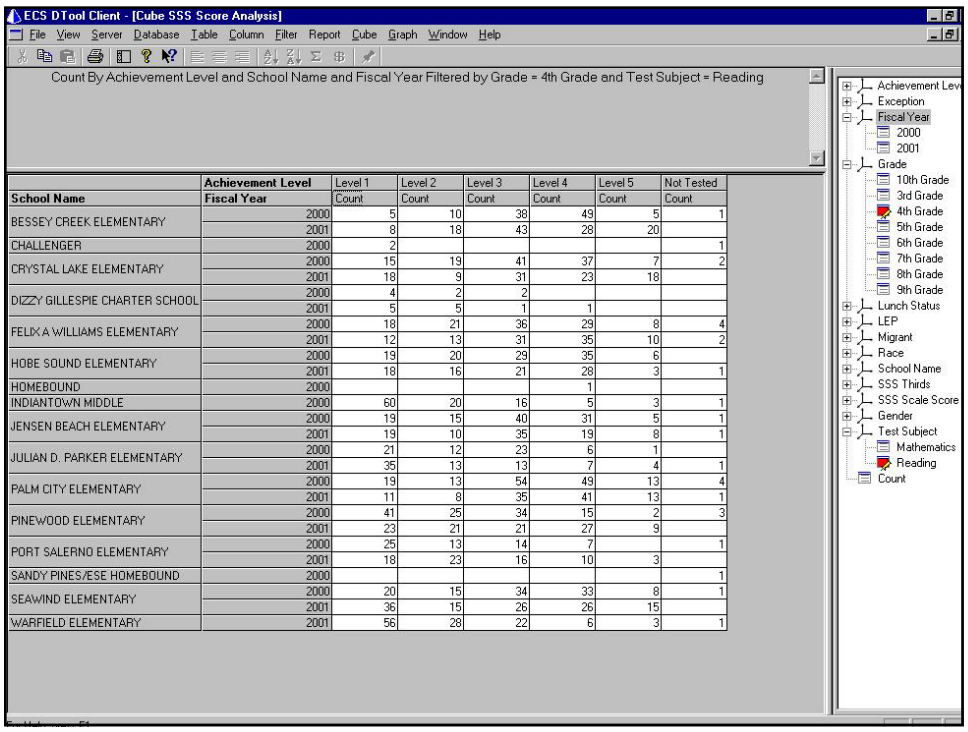

# OLAP Tools

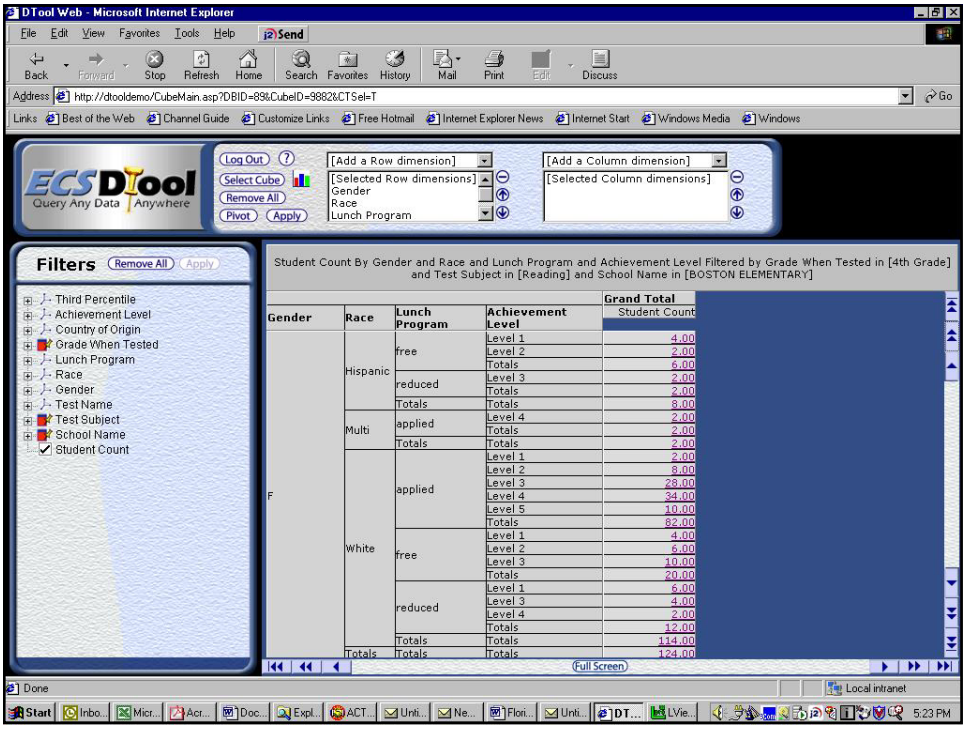
- Existing Tools: Crystal Decisions, Brio, Cognos
  - Choice of tables
  - Allow user to specify relationships
  - Use of filtering conditions
  - Construction of "cubes on the fly"
- Main Problems:
  - Ambiguous semantics of aggregations across tables, performance for multiple dimension cubes

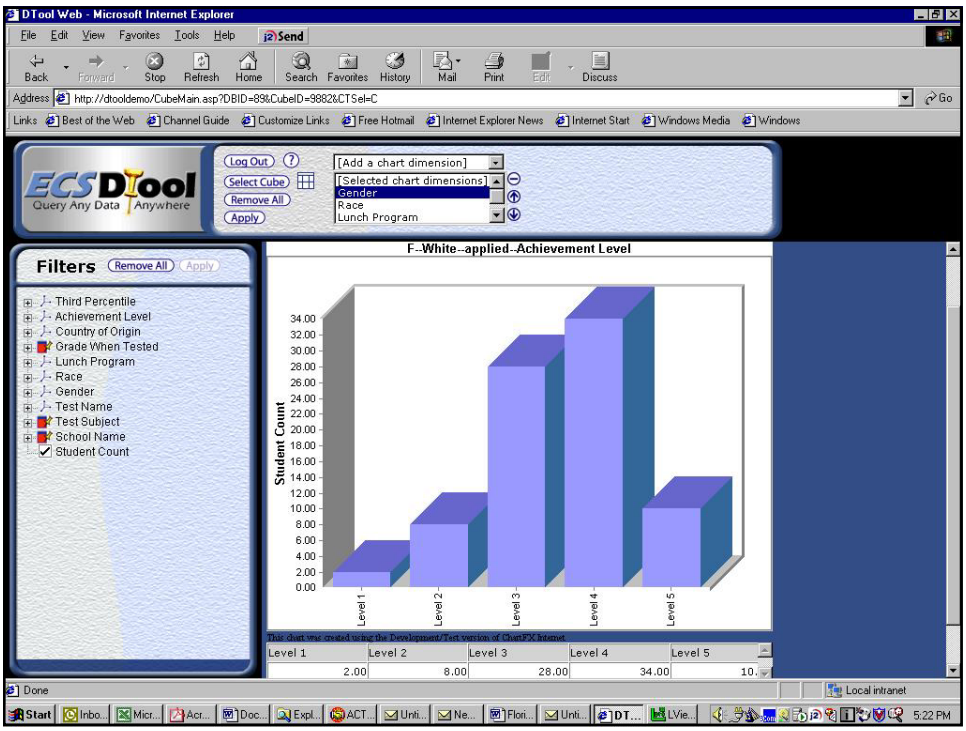
# A superior querying and OLAP Tool

- *ECS*DTool (from ECS Inc.)
  - Automatic detection and drawing of interrelation relationships
  - Automatic propagation of filtering conditions
  - Efficient Loading of "cubes on the fly"
  - Correct semantics of filters and aggregates across tables
- We have used many results from OLAP, DW and DM research to develop an intuitive point-and-click browsing and analysis OLAP tool
- More info: www.ecsdtool.com




DTool Web - Microsoft Internet Explorer
File Edit View Favorites Tools Help Send
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://dtooldemo/CubeMain.asp?DBID=89&CubeID=9882&CTSel=T
Links Best of the Web Channel Guide Customize Links Free Hotmail Internet Explorer News Internet Start Windows Media Windows
ECS DTool
Query Any Data Anywhere
Log Out
Select Cube
Remove All
Pivot
Apply
[Add a Row dimension]
[Selected Row dimensions]
Gender
Race
Lunch Program
[Add a Column dimension]
[Selected Column dimensions]
Filters
Remove All
Apply
Third Percentile
Achievement Level
Country of Origin
Grade When Tested
Lunch Program
Race
Gender
Test Name
Test Subject
School Name
Student Count
Student Count By Gender and Race and Lunch Program and Achievement Level Filtered by Grade When Tested in [4th Grade] and Test Subject in [Reading] and School Name in [BOSTON ELEMENTARY]
Full Screen
Done
Local intranet
Start
5:23 PM


DTool Web - Microsoft Internet Explorer
File Edit View Favorites Tools Help Send
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://dtooldemo/CubeMain.asp?DBID=89&CubeID=9882&CTSel=C
Links Best of the Web Channel Guide Customize Links Free Hotmail Internet Explorer News Internet Start Windows Media Windows
ECS DTool
Query Any Data Anywhere
Log Out
Select Cube
Remove All
Apply
[Add a chart dimension]
[Selected chart dimensions]
Gender
Race
Lunch Program
Filters
Remove All
Apply
Third Percentile
Achievement Level
Country of Origin
Grade When Tested
Lunch Program
Race
Gender
Test Name
Test Subject
School Name
Student Count
F--White--applied--Achievement Level
Student Count
34.00
32.00
30.00
28.00
26.00
24.00
22.00
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Level 1
Level 2
Level 3
Level 4
Level 5
This chart was created using the Development/Test version of ChartFX Internet
Level 1 Level 2 Level 3 Level 4 Level 5
2.00 8.00 28.00 34.00 10.
Done
Local intranet
Start
5:22 PM

# Recommendations

- Data Warehouses act as an excellent decision-support tool
- Using previously designed DW models lowers cost significantly
- Incorporating ETL tools (extract, transform and load) to automate data refresh
- Data Marts will allow a specific targeted analysis of current data by country or region etc.
- OLAP and Query tools help identify important trends for early action