AIR Forum 2005

MISSION: IMPROVE HIGHER EDUCATION

May 30, 2005

Datawarehousing for Institutional Research:

Experiences, Challenges and Goals

Leo Vélez-Ramos
University of Puerto Rico: Mayaguez Campus
Institutional Research and Planning Office

Monday, 11:10 a.m. - 11:50 a.m.

Agenda

- Background
- Where We Are
- Our Vision
- Data Warehouse
- Documentation
- Recommendations
- Conclusions

Background

- University of Puerto Rico – Mayaguez
  - Mid-size, public Institution
  - Degrees in
    - Engineering, Agriculture, Arts, Science, Business
  - 11,000 undergrads, 1000 grads
  - 900 faculty
  - 2000 staff
- Institutional Research Office
  - Established in 2001

Where We Are

- Campus data resides on three transactional systems (Open VMS system)
  - Student Information System (SIS)
  - Human Resources System (HRS)
  - Financial Resources System (FRS)
- Many years of historical data
- Reports must be requested to the Campus Computer Center
  - Custom programming (in COBOL) is needed to complete the report.

Our Vision

- To provide a data warehouse based system to fulfill ad-hoc reporting needs
  - Eliminate the need for custom programming
  - Easy to use
  - Available for everyone who needs it
- To provide other information not available through the warehouse
  - Data files with specific structures
  - Cross referencing of several tables
  - Special studies of interest

What is a Data Warehouse?

- A consolidated view of our enterprise data, optimized for reporting and analysis
- A "snapshot" of transaction data, NOT live
- Use data for analysis, NOT operational
  - "Transcripts are on the warehouse, but official transcripts are only available through SIS".
- Structure follows dimensional modeling techniques
  - Each business process enumerates relevant dimensions and facts.

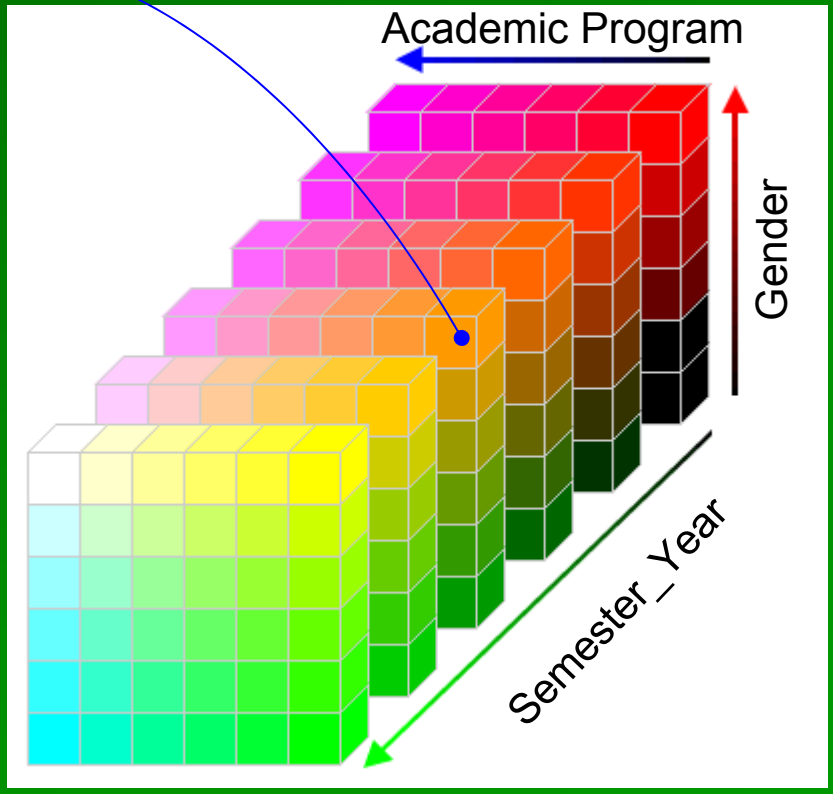
Dimensional Modeling: What is it?

- "A new name for an old technique"
  - Makes databases simple and understandable
  - Is a reflection of the manner in which a business process is viewed
- Data can be viewed as a cube
  - Many dimensions
  - Data can be "sliced" along any set of dimensions
    - Choosing one or more attributes
      - "Enrollment by gender and citizenship"
    - Applying constraints to any attributes
      - "for students in the school of business"

Dimensional Model of a Business: Student Enrollment

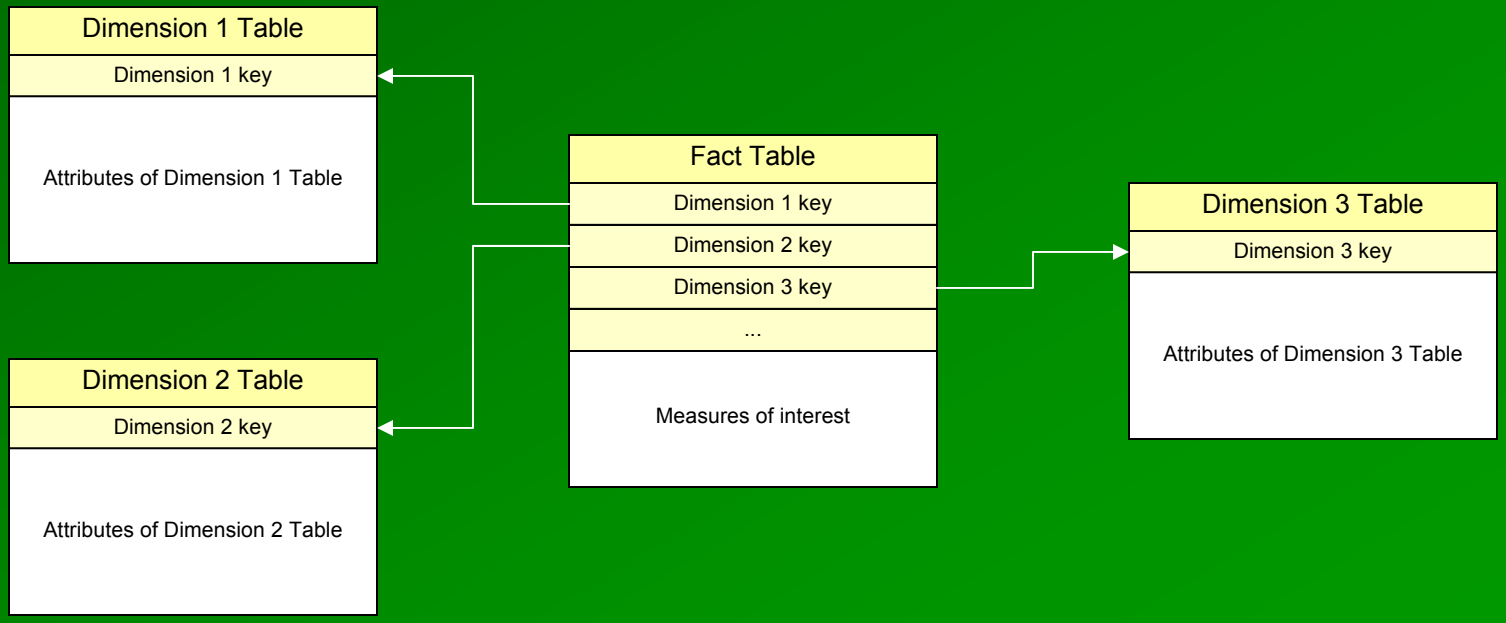
Each point in the cube contains measurements for a particular Combination of *academic program*, *gender* and *time*.

- Total Enrollment by
  - academic program
  - Gender
  - year and semester



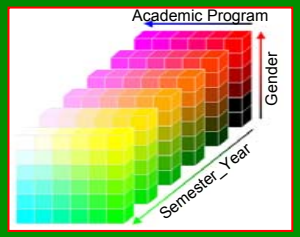
Tables in the Dimensional Model

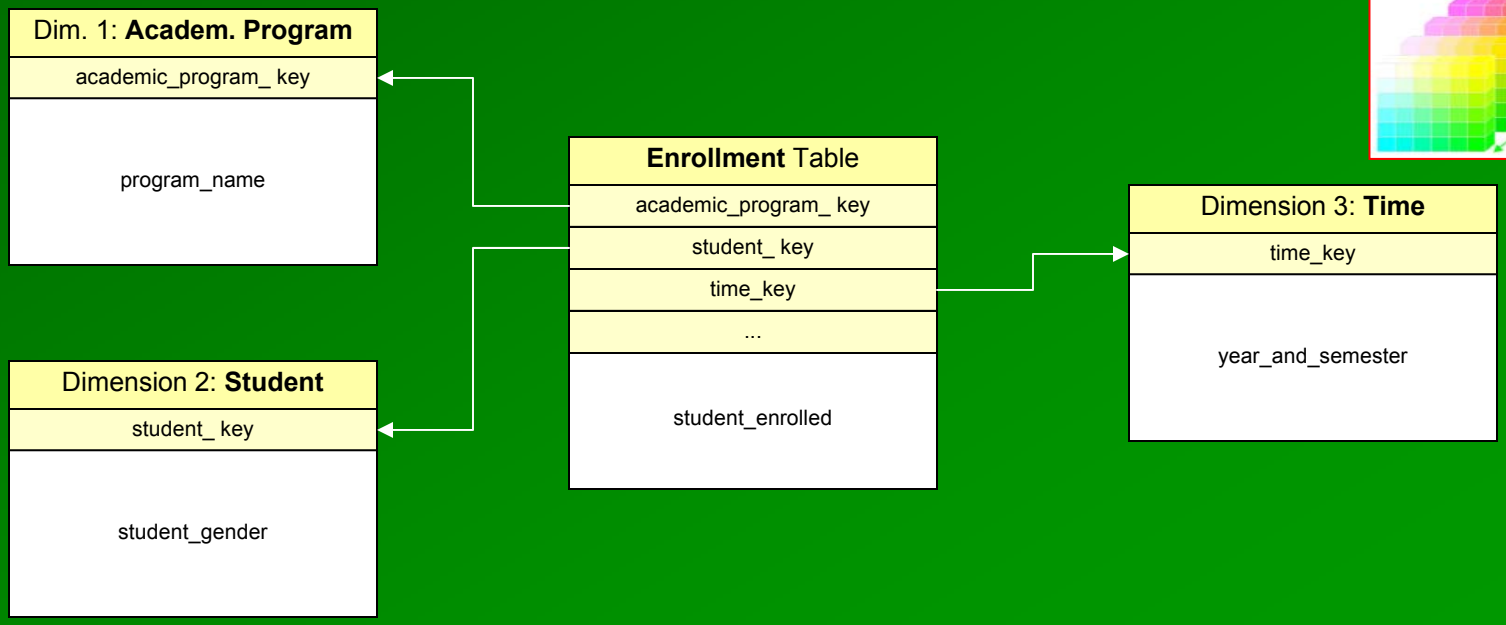
- All data is contained in two types of tables
  - Fact Tables (what you are measuring)
  - Dimension Tables (context for those measures)



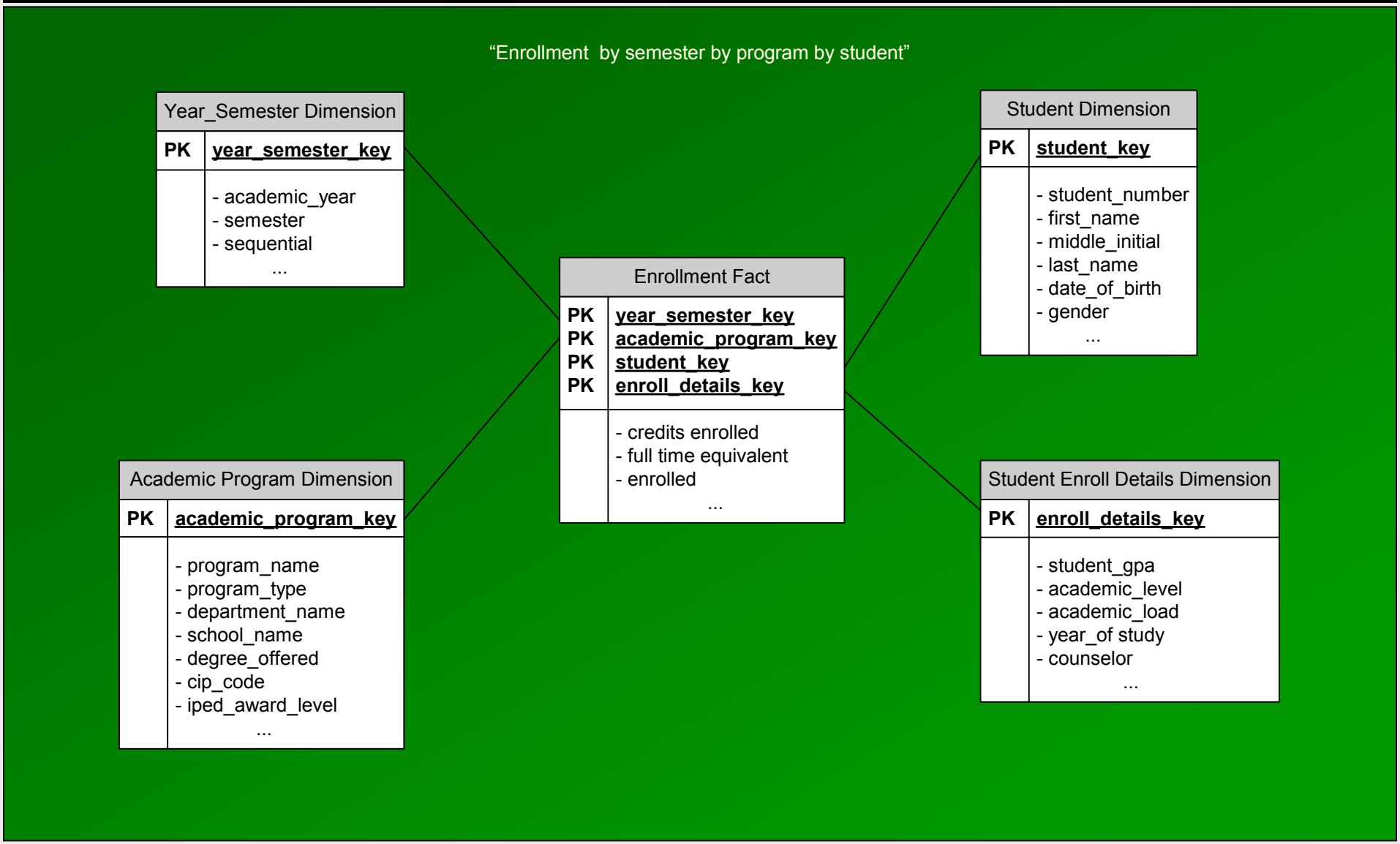
Tables in the Dimensional Model

- All data is contained in two types of tables
  - Fact Tables (what you are measuring)
  - Dimension Tables (context for those measures)

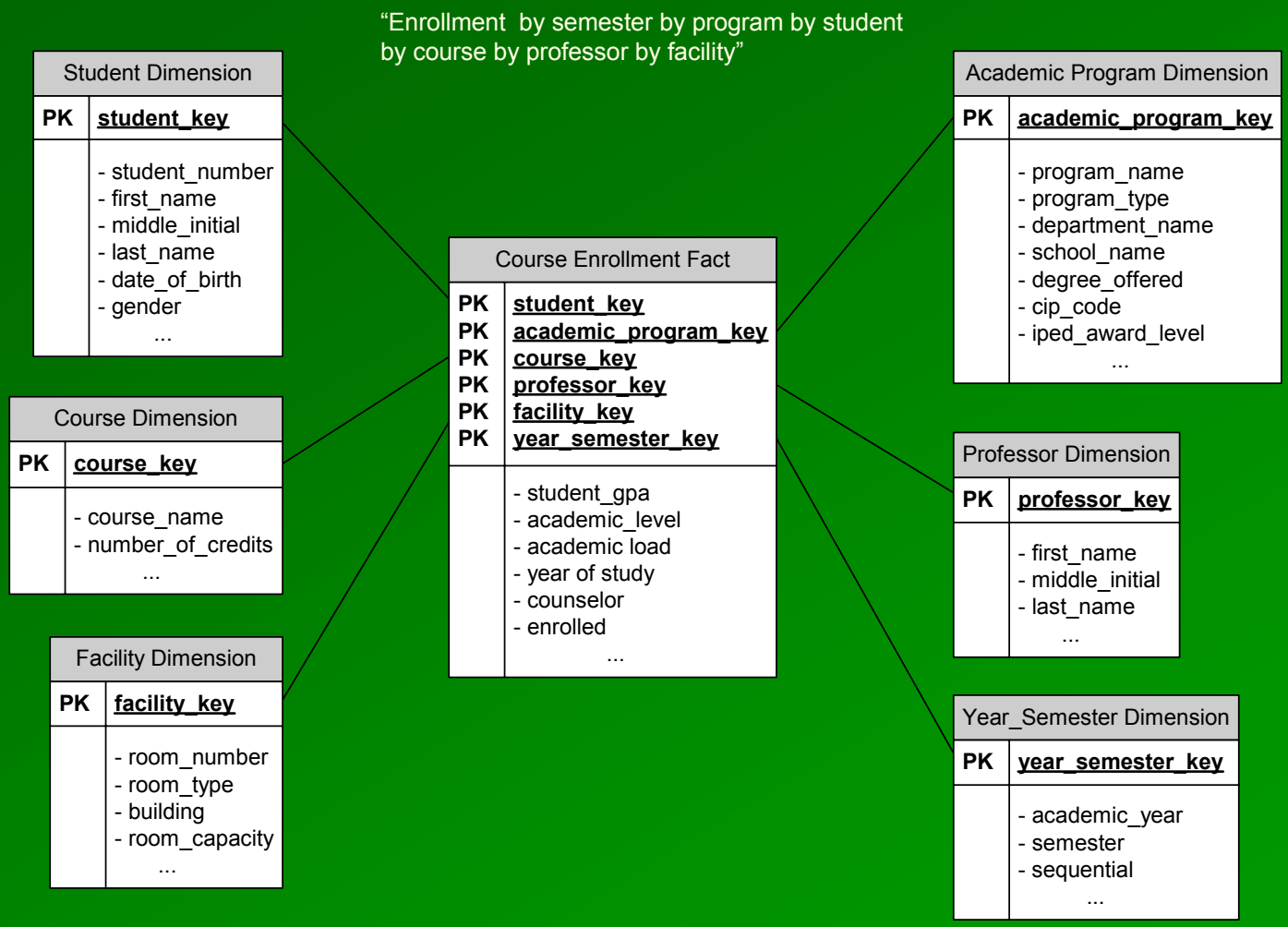




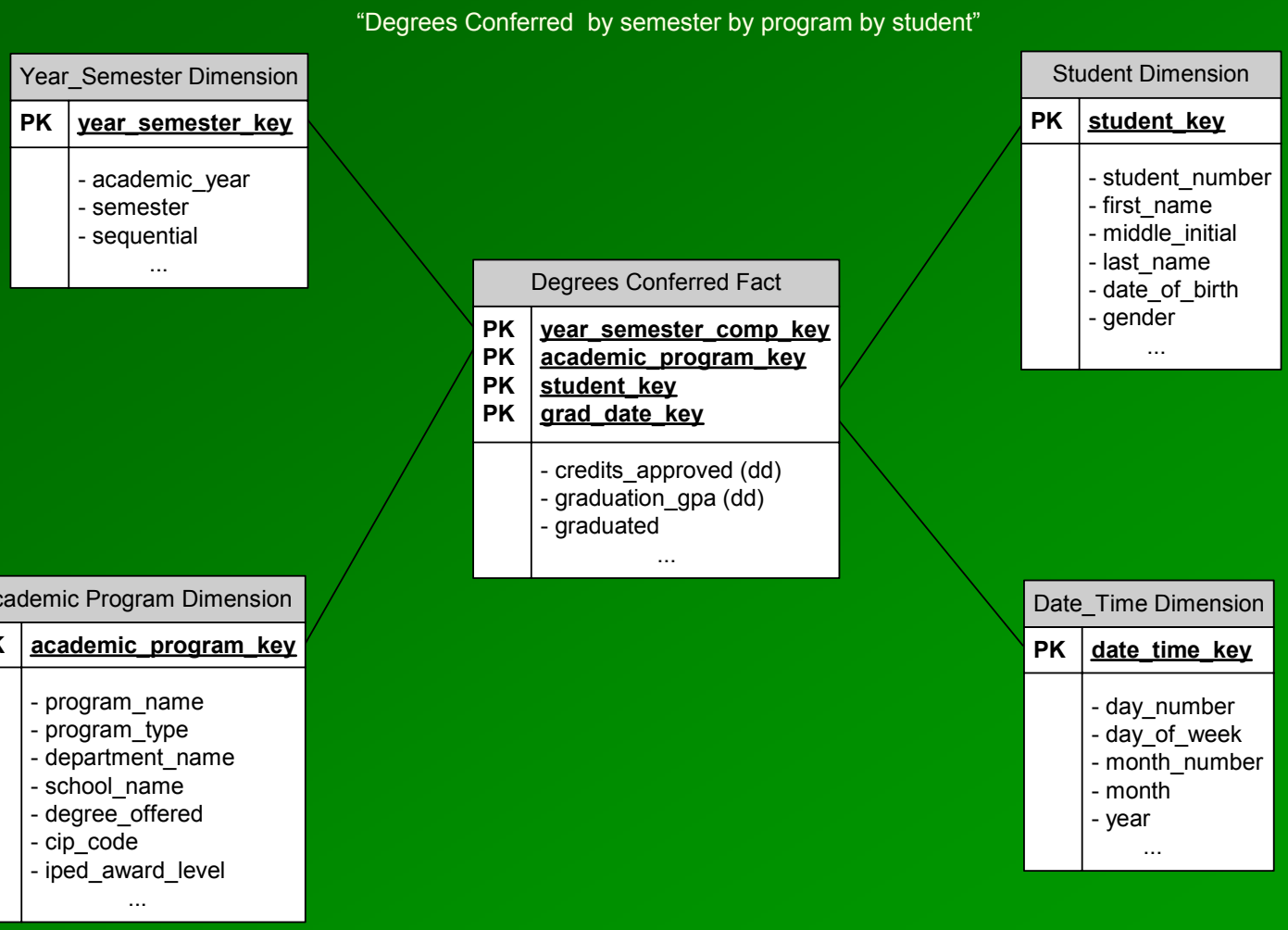
Student Enrollment Schema



Course Enrollment Schema



Degrees Conferred Schema



How Data Gets Into the Warehouse

- Data goes through a series of steps as it is moved to the warehouse.
  - Extraction from the legacy system
  - Data transformation (Cleansing)
  - Quality Assurance
  - Publishing
- Must be carried out periodically, in order to "refresh" the warehouse
  - Daily, weekly, once per semester

How Data Gets Into the Warehouse

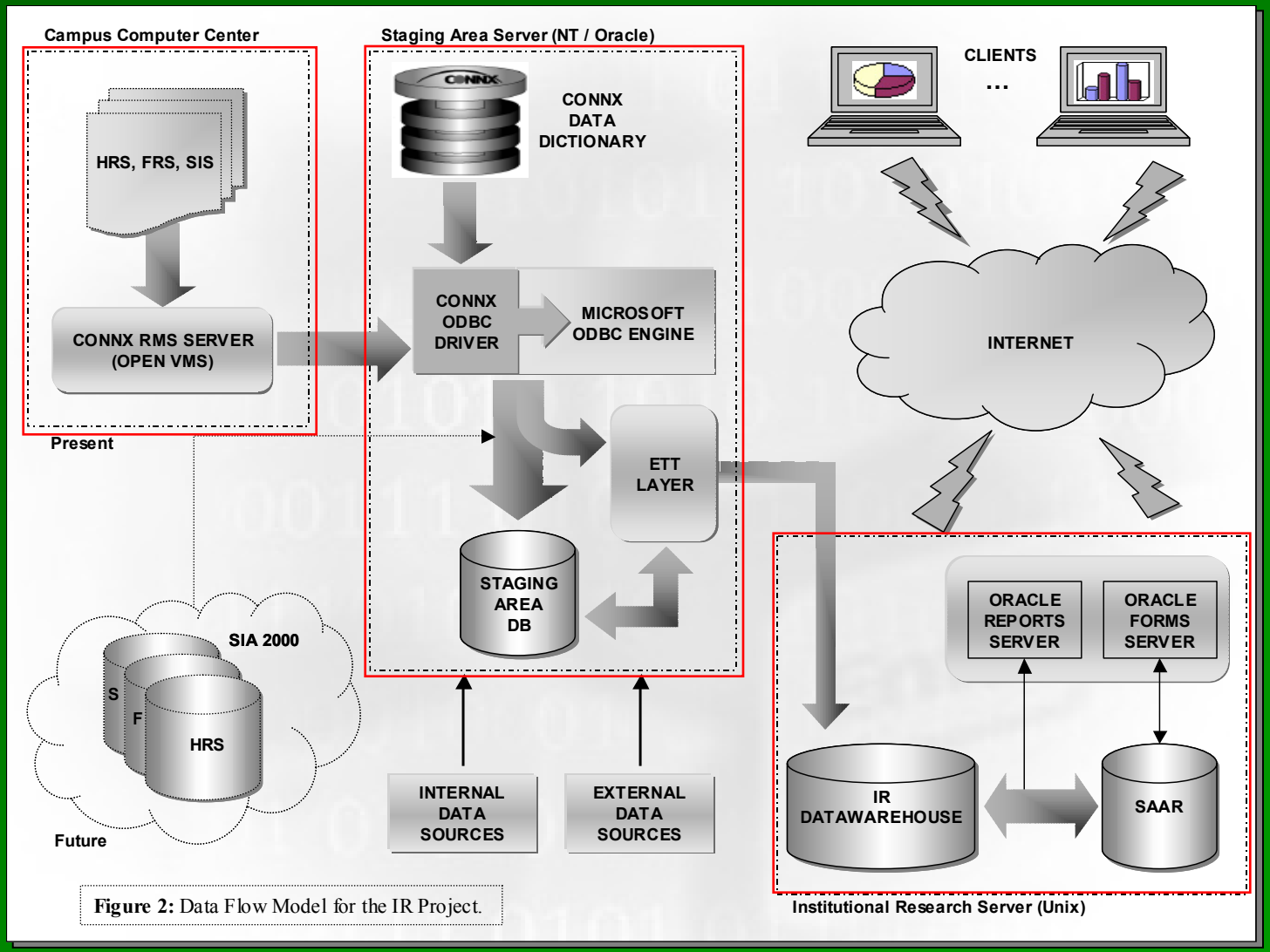
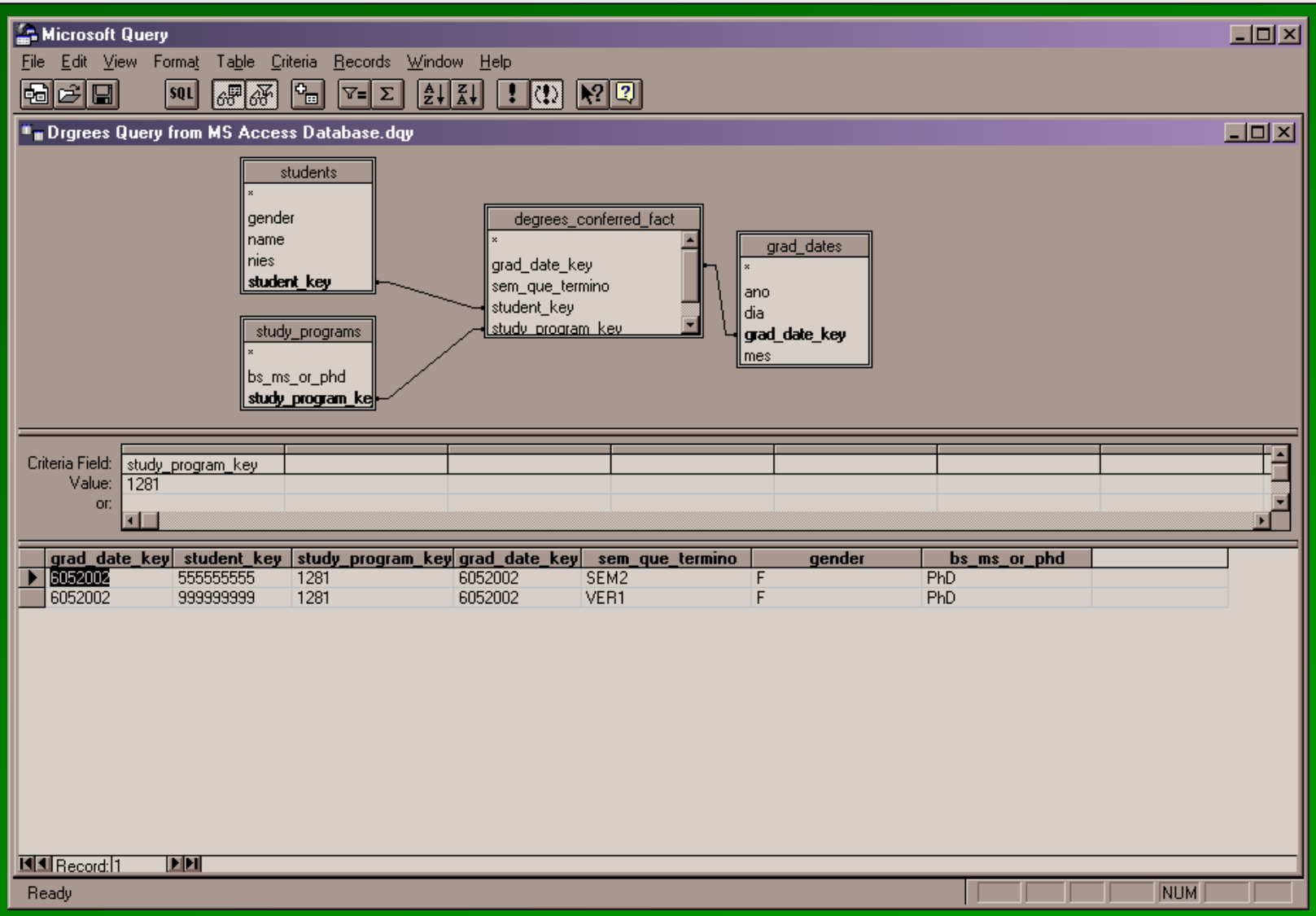


Figure 2: Data Flow Model for the IR Project.

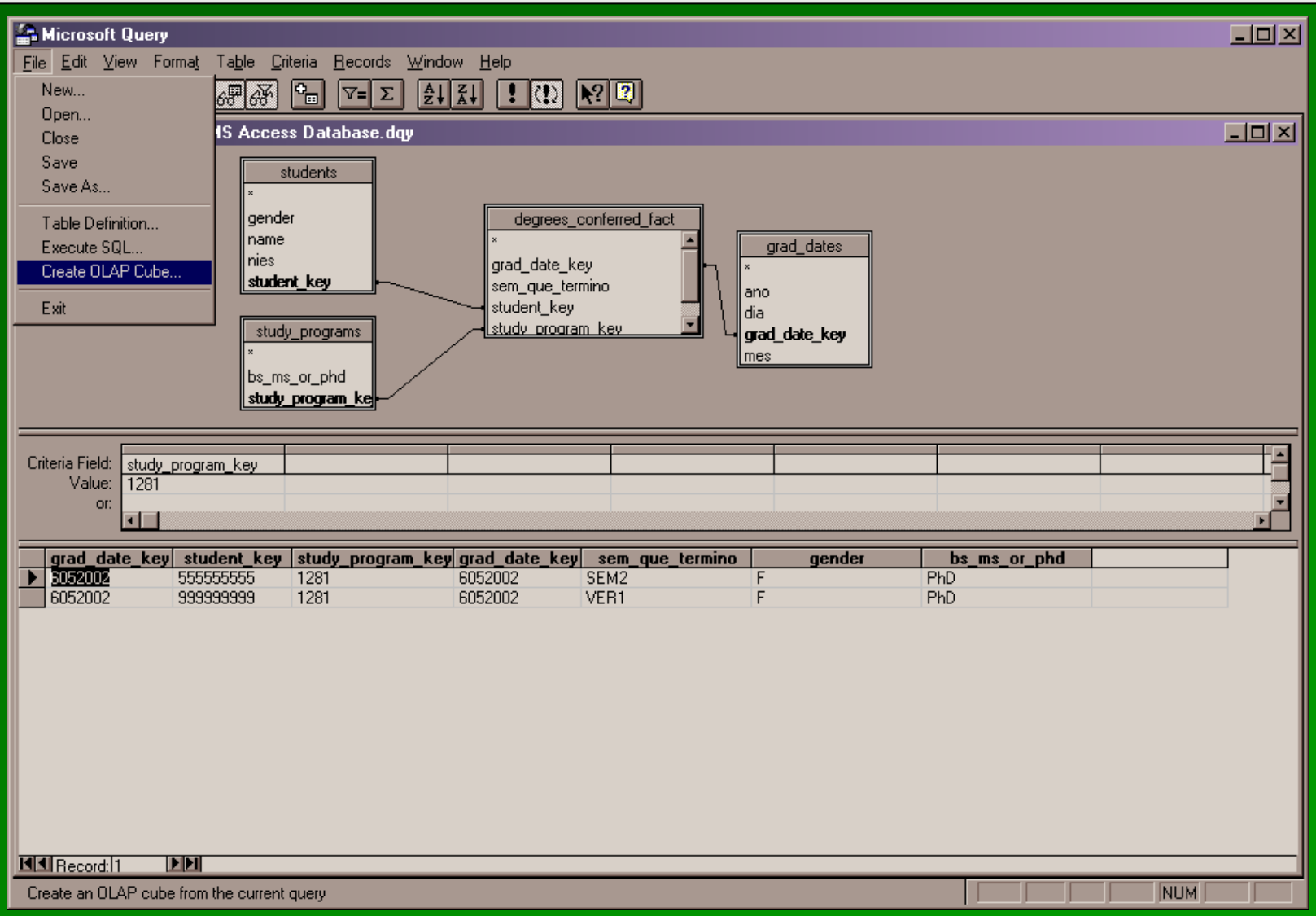
Getting Information from the Warehouse

- Use any tool that supports
  - Open Database Connectivity (ODBC)
    - Microsoft Access
    - Microsoft Excel
    - Microsoft Query
  - Direct queries to an Oracle DBMS
    - Many commercially available
- User must understand the structure
  - Documentation and/or training needed

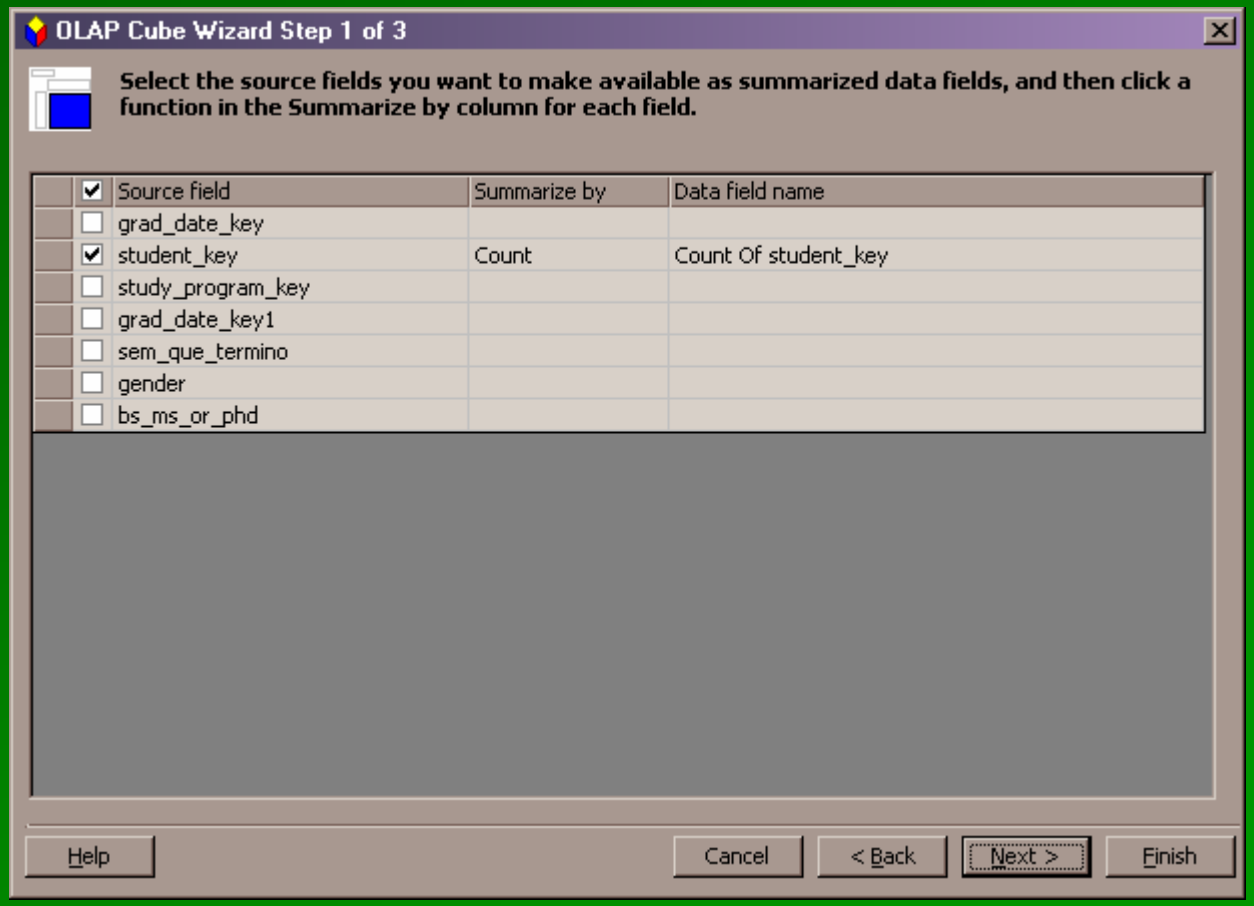
Extracting Information From a Schema



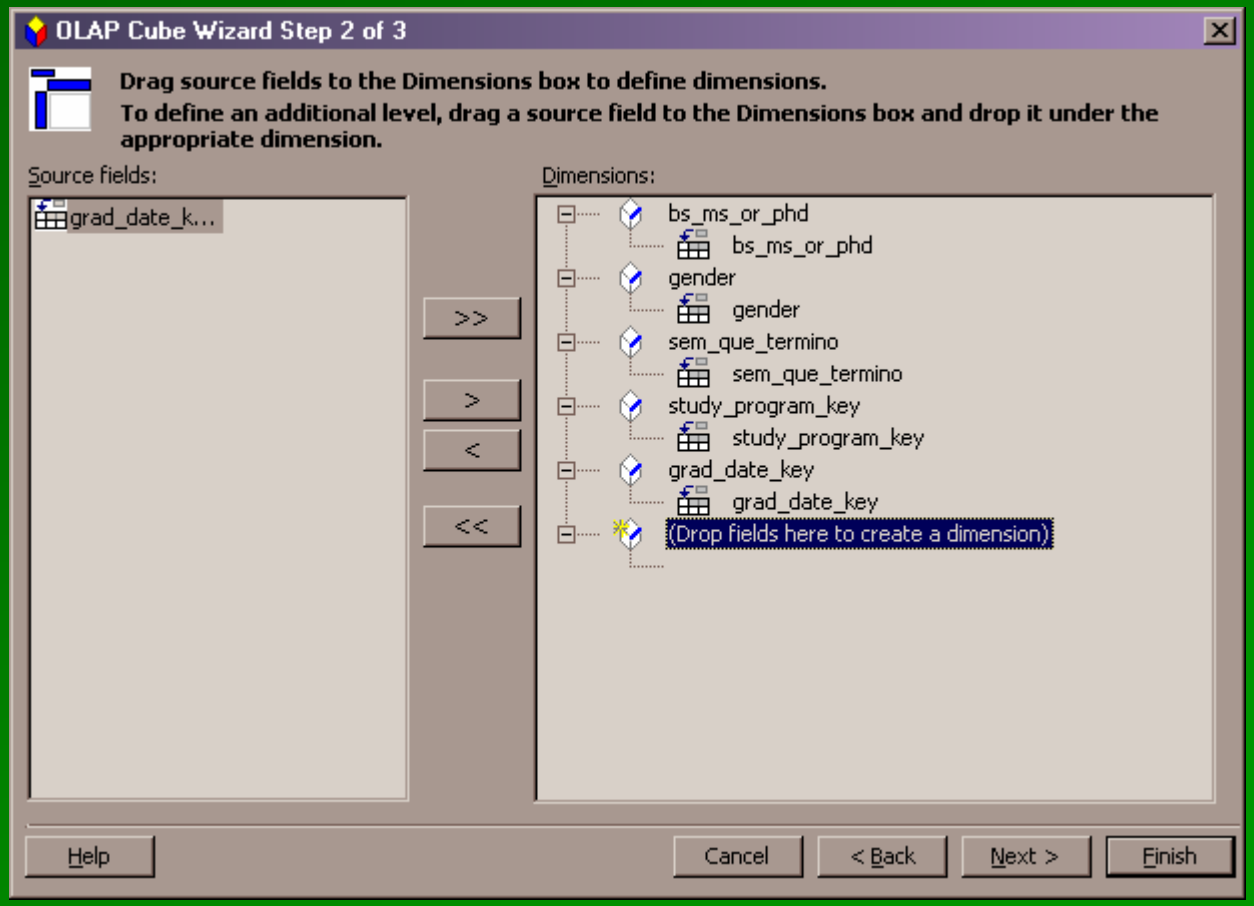
Extracting Information From a Schema



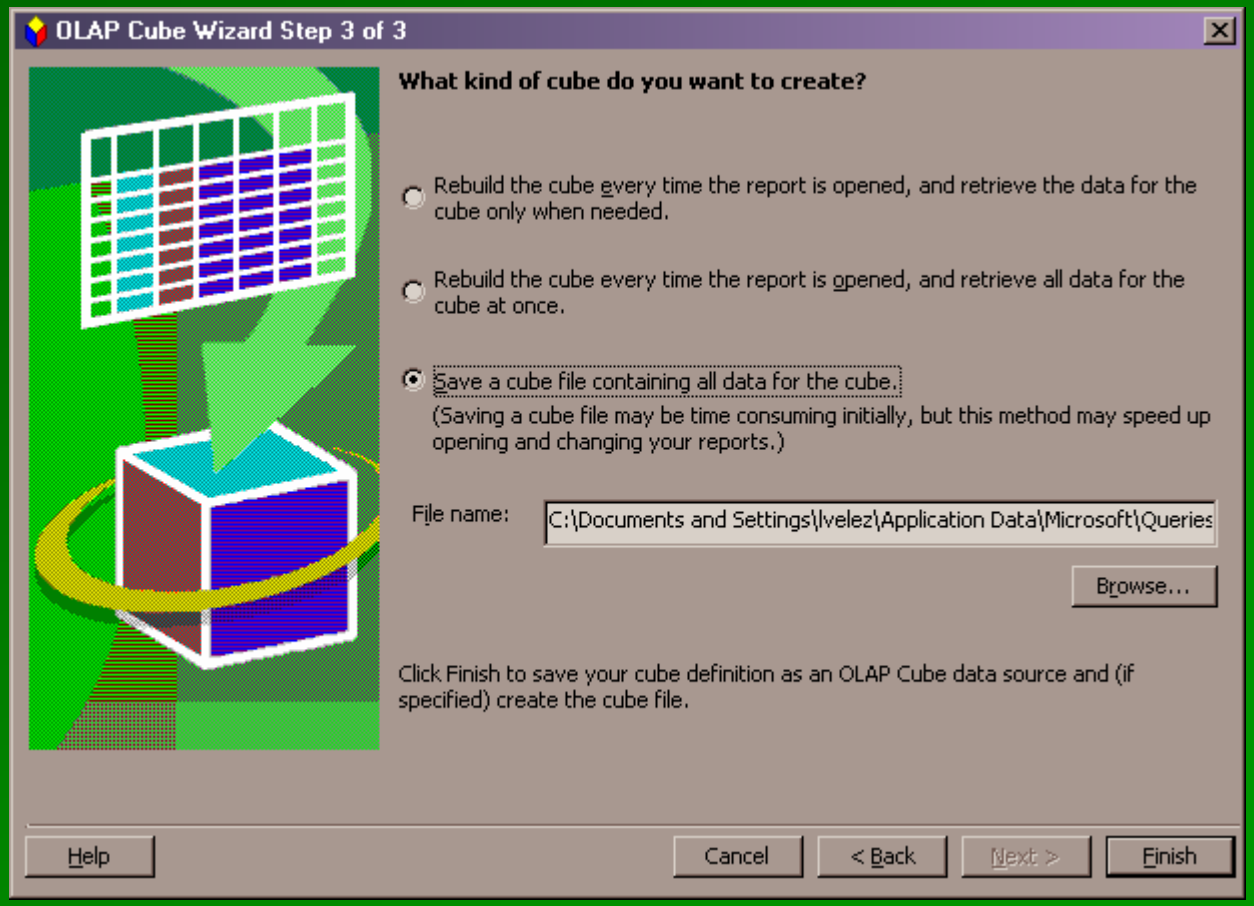
Extracting Information From a Schema



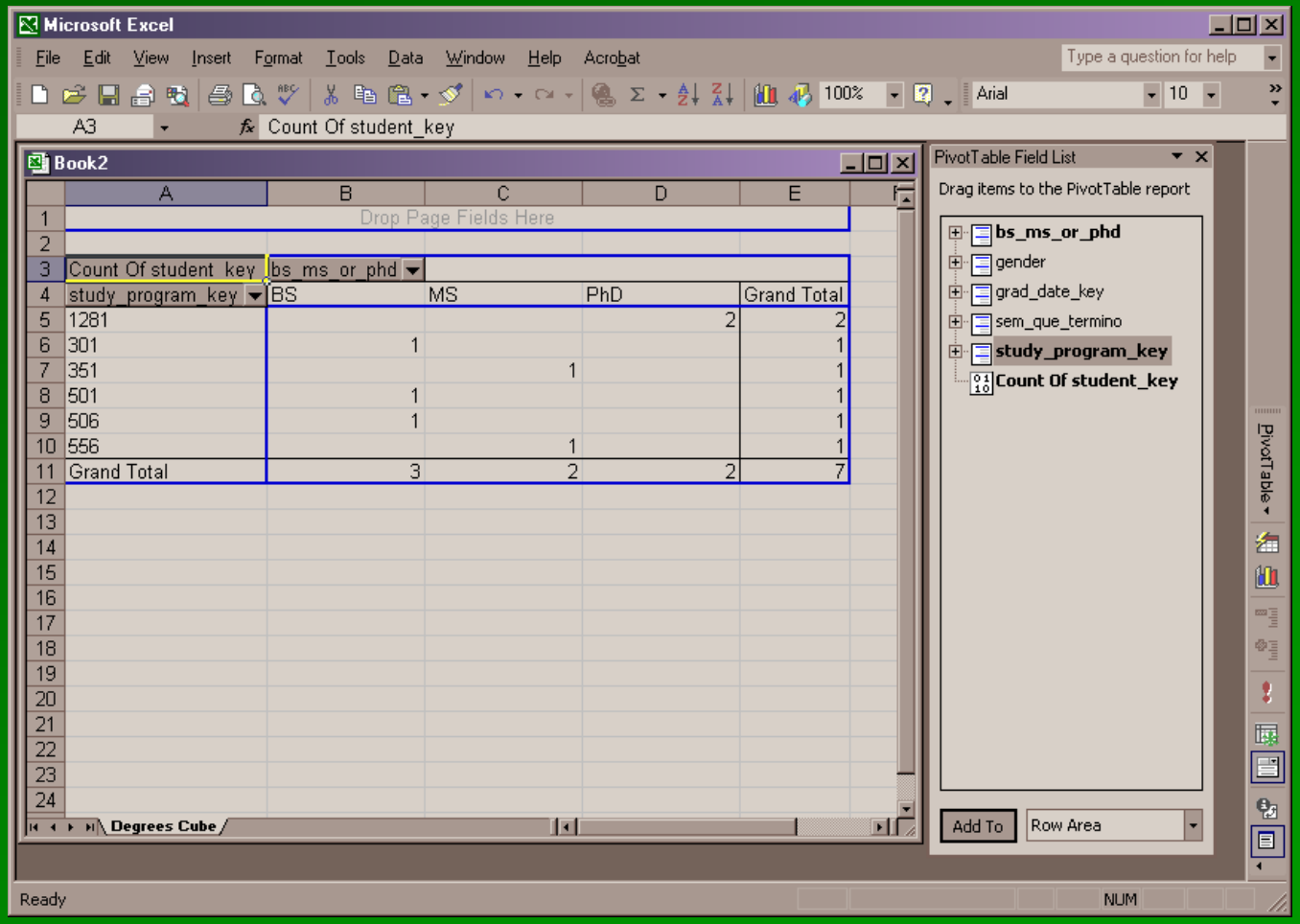
Extracting Information From a Schema



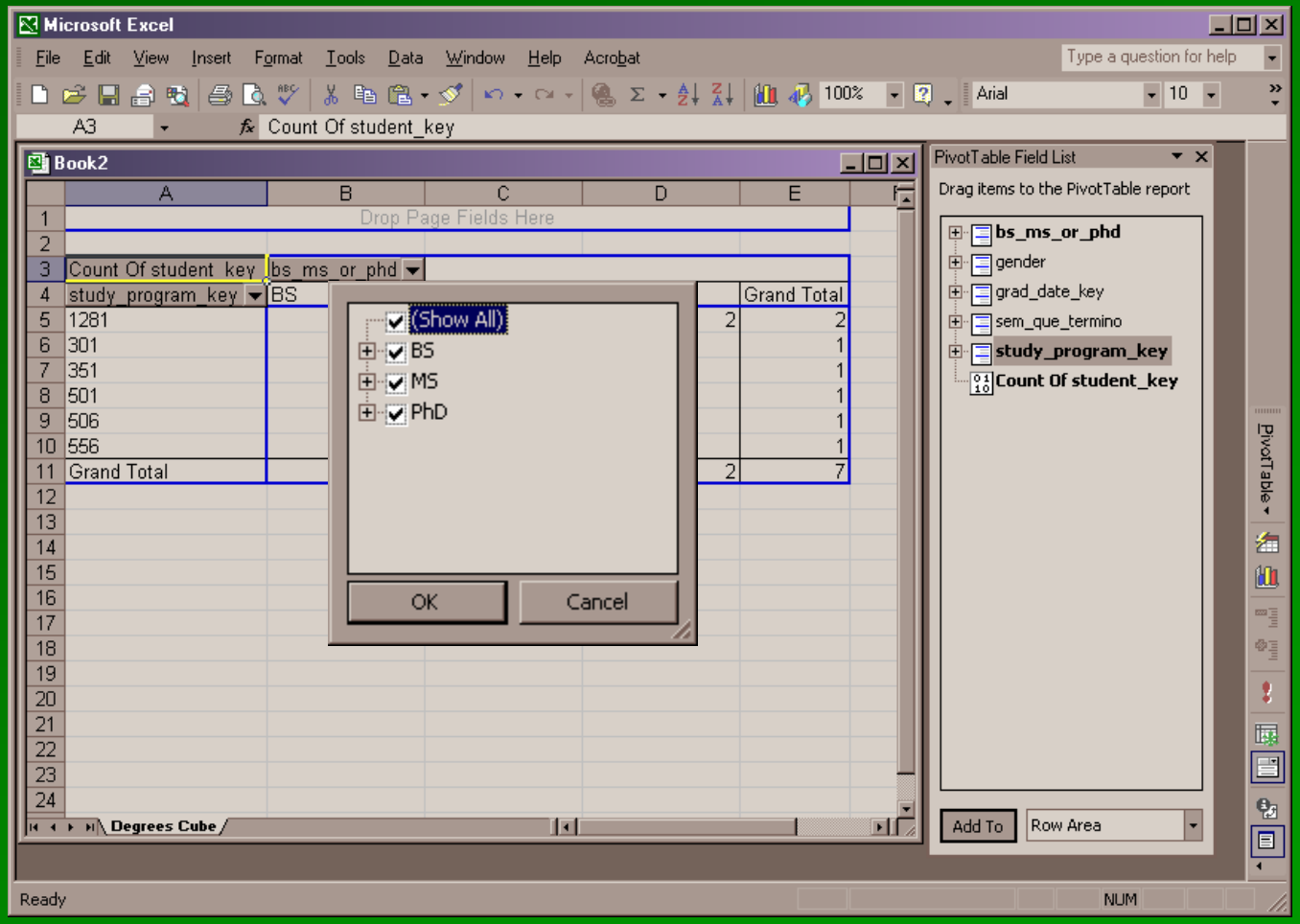
Extracting Information From a Schema



Extracting Information From a Schema



Extracting Information From a Schema



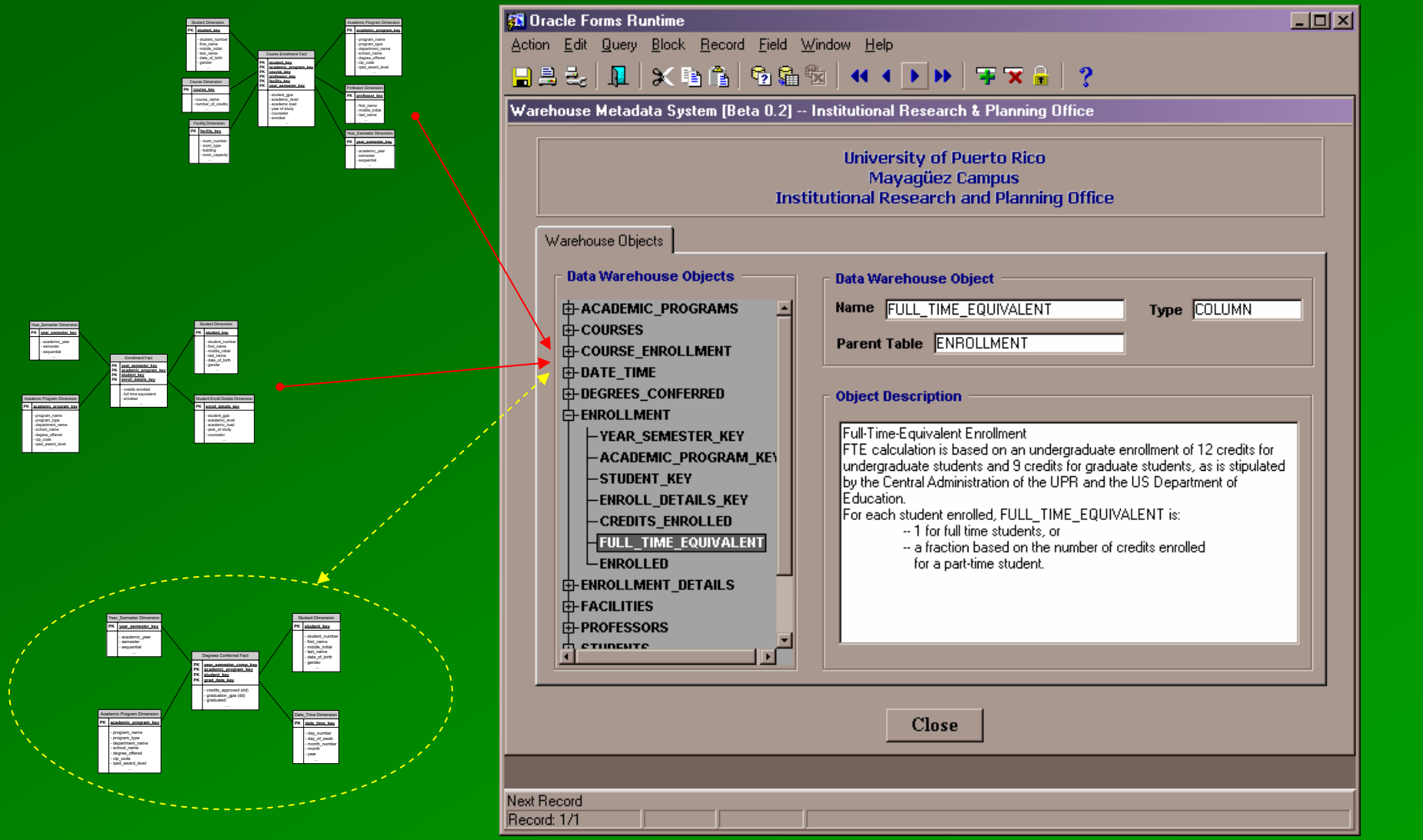
Warehouse Documentation

- Critical to the success of a DW project
- Missing or incomplete information increases usability problems
  - Users can become frustrated.
  - Will stop using the system if they can get their data by other means
  - Will speak negatively about the system
- Fiction: It's so easy that it doesn't need any documentation!

Warehouse Documentation Database

- Maintained by a custom system, currently being developed.
  - All warehouse tables and field descriptions are stored in a “documentation” database.
  - New warehouse schemas (facts/dimensions) are automatically loaded into the system.
  - Descriptions are then added to those new objects loaded.

Warehouse Documentation Database



Warehouse Documentation Database

- Advantages
  - Data is kept centralized
  - Dynamic reports can provide information in various formats if needed
  - Reports are updated automatically as data is modified or new information is added to the system

Tools we've Been Using

- Oracle DBMS (Currently on Version 8i)
- Oracle Developer Applications
  - Oracle Forms
  - Oracle Reports
- Data migration Tools
  - Connx Data Dictionary & ODBC Driver
- Data Analysis
  - SPSS
  - Minitab
  - MS Access / MS Excel / MS Query

What's been done

- Established data migration procedures for
  - SIS (Daily, Per Semester)
  - HRS (Weekly)
  - FRS (Daily)
- Developed several Datamarts (Schemas), based mostly on SIS
  - Enrollment (student / course level)
  - Grades
  - Conferred Degrees
  - Admissions (not completed)

What's been done

- Custom Queries (Cubes) and Reports
  - Created upon user request
  - Placed on our website for future users
- System for managing table and field documentation.
- Established a method for taking and processing user requests.
  - Previously done by IT people

The Future

- Establish the data warehouse as the authoritative source for all decision support data, thru
  - Institutional policies
  - Everyday use
- Design and implement other schemas
  - HRS and FRS related processes (Mostly)
    - Employee Analysis
    - Financial Analysis
  - Other student activities
    - Dropouts, Transfers, etc.
  - Cost of tuition

The Future

- Better tools
  - combine multiple datasources into a warehouse
  - reduce development time
  - provide friendlier user front-ends
    - Ease of use is critical to the acceptance of a tool
- Provide ways to access the DW over a web browser
  - Mostly for predefined reports
  - Security is important
- Improve the documentation system
- Provide online access to documentation

Recommendations

- Dimensions should be filled with as many descriptive attributes as possible
  - This greatly enhances the ability to “slice” through data
- Dimensions should contain both data codes as well as descriptions (readable text strings)
  - Example
    - `1201 : Biology`
    - `1204 : Chemistry`
    - `1205 : Mathematics`

Recommendations

- Use custom views to provide data access
  - Isolate fields (security)
  - Provide alternate names for a particular object
    - Good for handling multiple languages
- Use long field names
  - Example
    - `fte` vs `full_time_equivalent`

Conclusion

- A data warehouse is a continuous project
- A data warehouse is not just data, but also a set of tools to query, analyze and present information
- Develop the data warehouse incrementally
  - Visualize the warehouse as a series of small schemas
  - Build your warehouse, one schema at a time
- Sound documentation is critical to the success of a DW project

More Information on Datawarehousing

- The Kimball Group
  - http://www.kimballgroup.com
- DM Review Magazine
  - http://www.dmreview.com
- Datawarehousing
  - http://www.datawarehouse.com
  - http://www.datawarehousing.com
- Connx Data Solutions
  - http://www.connx.com

Questions

Thank You!

Leo O. Vélez-Ramos

Auxiliary Researcher

University of Puerto Rico - Mayaguez

lvelez@uprm.edu

http://oiip.uprm.edu/pres.html