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

Datawarehousing for Institutional Research: Experiences, Challenges and Goals.
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
Datawarehousing for Institutional Research: Experiences, Challenges and Goals.

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”
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)

<table>
<thead>
<tr>
<th>Dimension 1 Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 1 key</td>
</tr>
<tr>
<td>Attributes of Dimension 1 Table</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 2 Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 2 key</td>
</tr>
<tr>
<td>Attributes of Dimension 2 Table</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 3 Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 3 key</td>
</tr>
<tr>
<td>Attributes of Dimension 3 Table</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fact Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 1 key</td>
</tr>
<tr>
<td>Dimension 2 key</td>
</tr>
<tr>
<td>Dimension 3 key</td>
</tr>
<tr>
<td>Measures of interest</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 1 Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 1 key</td>
</tr>
<tr>
<td>Attributes of Dimension 1 Table</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 2 Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 2 key</td>
</tr>
<tr>
<td>Attributes of Dimension 2 Table</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 3 Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension 3 key</td>
</tr>
<tr>
<td>Attributes of Dimension 3 Table</td>
</tr>
</tbody>
</table>
All data is contained in two types of tables
- Fact Tables (what you are measuring)
- Dimension Tables (context for those measures)
Student Enrollment Schema

“Enrollment by semester by program by student”

<table>
<thead>
<tr>
<th>Year_Semester Dimension</th>
<th>PK</th>
<th>year_semester_key</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PK</td>
<td>academic_year</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>semester</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>sequential</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>enroll_details_key</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>credits enrolled</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>full time equivalent</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>enrolled</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Academic Program Dimension</th>
<th>PK</th>
<th>academic_program_key</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PK</td>
<td>program_name</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>program_type</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>department_name</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>school_name</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>degree_offered</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>cip_code</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>iped_award_level</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>student_gpa</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>academic_level</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>academic_load</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>year_of_study</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>counselor</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enrollment Fact</th>
<th>PK</th>
<th>year_semester_key</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PK</td>
<td>academic_program_key</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>student_key</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>enroll_details_key</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>credits enrolled</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>full time equivalent</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>enrolled</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Dimension</th>
<th>PK</th>
<th>student_key</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PK</td>
<td>student_number</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>first_name</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>middle_initial</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>last_name</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>date_of_birth</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>gender</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Enroll Details Dimension</th>
<th>PK</th>
<th>enroll_details_key</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PK</td>
<td>student_gpa</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>academic_level</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>academic_load</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>year_of_study</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>counselor</td>
</tr>
</tbody>
</table>

May 30, 2005

Datawarehousing for Institutional Research: Experiences, Challenges and Goals.

11
Course Enrollment Schema

Course Enrollment Fact

PK
- student_key
- academic_program_key
- course_key
- professor_key
- facility_key
- year_semester_key

“Enrollment by semester by program by student by course by professor by facility”

Student Dimension
PK
- student_key
- student_number
- first_name
- middle_initial
- last_name
- date_of_birth
- gender
...

Course Dimension
PK
- course_key
- course_name
- number_of_credits
...

Facility Dimension
PK
- facility_key
- room_number
- room_type
- building
- room_capacity
...

Academic Program Dimension
PK
- academic_program_key
- program_name
- program_type
- department_name
- school_name
- degree_offered
- cip_code
- iped_award_level
...

Professor Dimension
PK
- professor_key
- first_name
- middle_initial
- last_name
...

Year_Semester Dimension
PK
- year_semester_key
- academic_year
- semester
- sequential
...
Degrees Conferred by semester by program by student

**Table: Degrees Conferred Schema**

**Year_Semester Dimension**
- PK year_semester_key
- academic_year
- semester
- sequential

**Student Dimension**
- PK student_key
- student_number
- first_name
- middle_initial
- last_name
- date_of_birth
- gender

**Academic Program Dimension**
- PK academic_program_key
- program_name
- program_type
- department_name
- school_name
- degree_offered
- cip_code
- iped_award_level

**Degrees Conferred Fact**
- PK year_semester_comp_key
- academic_program_key
- student_key
- grad_date_key
- credits_approved (dd)
- graduation_gpa (dd)
- graduated

**Date_Time Dimension**
- PK date_time_key
- day_number
- day_of_week
- month_number
- month
- year

---

May 30, 2005  
Datawarehousing for Institutional Research:  
Experiences, Challenges and Goals.  13
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

### OLAP Cube Wizard Step 1 of 3

Select the source fields you want to make available as summarized data fields, and then click a function in the Summarize by column for each field.

<table>
<thead>
<tr>
<th>Source field</th>
<th>Summarize by</th>
<th>Data field name</th>
</tr>
</thead>
<tbody>
<tr>
<td>grad_date_key</td>
<td></td>
<td></td>
</tr>
<tr>
<td>student_key</td>
<td>Count</td>
<td>Count Of student_key</td>
</tr>
<tr>
<td>study_program_key</td>
<td></td>
<td></td>
</tr>
<tr>
<td>grad_date_key1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sem_que_termino</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bs_ms_or_phd</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Insert diagram and table here]
Extracting Information From a Schema

OLAP Cube Wizard Step 2 of 3

Drag source fields to the Dimensions box to define dimensions.
To define an additional level, drag a source field to the Dimensions box and drop it under the appropriate dimension.

Source fields:
- grad_date_k...

Dimensions:
- bs_ms_or_phd
  - bs_ms_or_phd
- gender
  - gender
- sem_que_termino
  - sem_que_termino
- study_program_key
  - study_program_key
- grad_date_key
  - grad_date_key
(Drop fields here to create a dimension)
Extracting Information From a Schema

What kind of cube do you want to create?

- Rebuild the cube every time the report is opened, and retrieve the data for the cube only when needed.
- Rebuild the cube every time the report is opened, and retrieve all data for the cube at once.
- Save a cube file containing all data for the cube.
  (Saving a cube file may be time consuming initially, but this method may speed up opening and changing your reports.)

File name: `C:\Documents and Settings\velez\Application Data\Microsoft\Queries`

Click Finish to save your cube definition as an OLAP Cube data source and (if specified) create the cube file.
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.
Datawarehousing for Institutional Research: Experiences, Challenges and Goals.
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