Introduction to UCDW Star Schemas and Data Marts

Data Infrastructure IRAP Training

3/20/2017
Agenda

- Introduction to UCDW
  - The three layered architecture
  - Star schemas and data marts
  - Differences – star schema & data mart
- Facts & characteristics
- Dimensions & characteristics
- UCDW conformed dimensions
- Slowly changing dimensions (SCD)
- UCDW Naming Conventions
- UCDW schemas and contents
- Live demo using DB Visualizer
- Questions & Answers
Introduction to UCDW

- Enterprise data warehouse
- 3 distinct environments
- Data sources
- Data load process
- Long term strategy
- Technology
- Server–schema–table or view–columns
- Connecting to UCDW
UCDW Environments

1. Development
   DWD2
2. Quality Assurance
   DWP3
3. Production
   DWP2

Extract–Transform–Load
The 3-Layered Architecture

- **Input data parking lot**
- **Cooking Area**
- **Reporting & Analytics**

1. **Staging Layer**
2. **Base Layer**
3. **BI Layer**

The diagram illustrates the flow of data from input data to reporting & analytics, with ETL processes at each layer.
Current UCDW Infrastructure

Stage

Base

BI

Enterprise Data Warehouse

Data Marts

Input

XML
What is a Star-Schema?

- Simplest form of a dimensional Model
- Diagram resembles a star
- One or more fact tables referenced by a number of dimensional tables
- Data is organized into fact and dimensions
Dimensional modeling

- Easier for business users to understand
- Query performance
- Symmetrical structure
- Each dimension is an entry point into the fact table
- Extensible to accommodate data changes
What is a Data Mart?

- Specific content for specific needs
- Subsets of data warehouse – holds one subject area
- Tactical
- Improved end-user response times
- IRAP’s use of data marts
Differences – Star vs. Data Mart

- Holds multiple subject areas
- Holds very detailed information
- Works to integrate all data sources
- Uses a dimensional model

- Only one subject area
- More high-level data
- Concentrates on integrating data within the same subject area
- Extracted from dimensional model
- Flat table structure

UCDW Data Warehouse

UCDW Data Mart
Facts

- Contains performance measurements, metrics or facts – quantitative data
- Located at center of a star schema
- Defined by its grain
- UCDW facts – periodic snapshots
- UCDW Examples
  - Enrollment fact – Enrollment
  - Degree Awarded fact – Degree
  - Application fact – Undergraduate Admissions
Fact Table Characteristics

- Fact table contains measures that are typically additive and are chronologically organized in the form of time series. For e.g. in Student Enrollment
  - Units attempted in the 3WK of fall 2016
- Measures are often stored in the fact table as a series of periodic snapshots. Examples:
  - Enrollment measures for fall end-of-term 2016
- Each fact row is an intersection of all dimension records.
- Each fact table has a specific grain: the lowest level at which the data is captured
Dimensions

- Descriptive fields
- Power of data warehouse
- Typically textual fields
- Act as filters or labels
- Includes codes and abbreviations

UCDW Examples
- Student Enrollment – Student Level
- Undergraduate Admission – Applicant Level
- Conformed – Campus–College–Major (CCM)
Attributes related to the measures are logically grouped in a number of dimensions. Examples:
- Student attributes (e.g., name, gender, etc.) are stored in the Student dimension
- Data about majors are stored in the CCM dimension
- Term data is stored in the Term dimension

Dimension attributes provide context to each of the measures that are stored in the fact table:
- Student Level: provides a context to the count of students (Graduate vs. Undergraduate)
### Table 1a: Enrollment by Campus, Level, and Gender: General Campus and Health Sciences Combined

<table>
<thead>
<tr>
<th></th>
<th>Fall 2011</th>
<th></th>
<th>Fall 2012</th>
<th></th>
<th>One-year change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Total</td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>San Diego</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergraduate</td>
<td>14,320</td>
<td>15,004</td>
<td>29,324</td>
<td>13,812</td>
<td>15,247</td>
</tr>
<tr>
<td>Lower Division</td>
<td>11,646</td>
<td>11,400</td>
<td>23,046</td>
<td>11,202</td>
<td>11,474</td>
</tr>
<tr>
<td>Upper Division</td>
<td>3,595</td>
<td>3,059</td>
<td>6,654</td>
<td>3,792</td>
<td>3,599</td>
</tr>
<tr>
<td>Graduate</td>
<td>8,051</td>
<td>8,341</td>
<td>16,392</td>
<td>7,410</td>
<td>7,875</td>
</tr>
<tr>
<td>1st Stage</td>
<td>2,319</td>
<td>3,228</td>
<td>5,547</td>
<td>2,242</td>
<td>3,376</td>
</tr>
<tr>
<td>2nd Stage</td>
<td>1,897</td>
<td>2,593</td>
<td>4,490</td>
<td>1,848</td>
<td>2,728</td>
</tr>
<tr>
<td>Resident</td>
<td>422</td>
<td>635</td>
<td>1,057</td>
<td>394</td>
<td>648</td>
</tr>
<tr>
<td></td>
<td>355</td>
<td>376</td>
<td>731</td>
<td>368</td>
<td>397</td>
</tr>
</tbody>
</table>
UCDW Conformed Dimensions

- Dimensions shared across the enterprise
  - STUDENT_D
  - AGE_BAND_D
  - CAMPUS_COLLEGE_MAJOR_D
  - ACADEMIC_DEGREE_D
  - CAMPUS_LOCATION_D
  - COUNTRY_STATE_COUNTY_D
  - IPEDS_ETHNIC_CODE_D
  - UC_ETHNIC_CODE_LEVEL_1
  - UC_ETHNIC_CODE_LEVEL_2
UCDW currently uses these two concepts:

- **SCD Type 1** – No history of changes is maintained. Only the latest/most current value is available. Used when old values have no business significance.
  - Example – If a student’s birth date changes on the enrollment file, the STUDENT_D row is simply updated with the new value.

- **SCD Type 2** – With a change within an attribute of a row, a new row is inserted into the dimension table to reflect the new attribute value.
  - Example – If a student’s last name changes, a new row is created in STUDENT_D.
**SCD Type 1:** Student’s date of birth changed from 1972–07–05 to 1975–07–05

<table>
<thead>
<tr>
<th>STUD_LOC_CMP_CD</th>
<th>STUD_ID</th>
<th>STUD_CUR.ACTV_FL</th>
<th>STUD_BEG_EFF_DT</th>
<th>STUD_END_EFF_DT</th>
<th>STUD_DT_OF_BTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>15084017</td>
<td>Y</td>
<td>2002-06-30</td>
<td>9999-12-31</td>
<td>1975-07-05</td>
</tr>
</tbody>
</table>

**SCD Type 2:** Student’s last name changed from Nelson to Shaw on June 30th 2002

<table>
<thead>
<tr>
<th>STUD_LOC_CMP_CD</th>
<th>STUD_ID</th>
<th>STUD_CUR.ACTV_FL</th>
<th>STUD_BEG_EFF_DT</th>
<th>STUD_END_EFF_DT</th>
<th>STUD_LST_NAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>15084017</td>
<td>N</td>
<td>2000-11-08</td>
<td>2002-06-29</td>
<td>NELSON</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STUD_LOC_CMP_CD</th>
<th>STUD_ID</th>
<th>STUD_CUR.ACTV_FL</th>
<th>STUD_BEG_EFF_DT</th>
<th>STUD_END_EFF_DT</th>
<th>STUD_LST_NAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>15084017</td>
<td>Y</td>
<td>2002-06-30</td>
<td>9999-12-31</td>
<td>SHAW</td>
</tr>
</tbody>
</table>

Most recent record
UCDW Naming Conventions

- Dimension tables have a suffix of _D
  - Examples: STUDENT_D, STUDENTLEVEL_D
- Fact tables have a suffix of _F
  - Examples: ENROLLMENT_F, DEGREE_AWARDED_F
- Views are database objects built on top of one or more tables to aggregate data or simplify usage. They have a suffix of _V (view) or _M (Materialized Query Table or MQT)
  - Examples: ENROLLMENT_HEAD_COUNT_M, IPEDS_ETHNIC_CODE_V
UCDW Schemas of Interest

- **GAD_BI**
  - Star – Graduate Admissions
- **IRAP_BI**
  - Miscellaneous Data Marts
- **DSS_BI**
  - Star – Payroll
- **SP_BI**
  - Star – Contracts & Grants
- **STUD_BI**
  - Stars – Student Enrollment, Course Enrollment, Degree and Financial Aid
- **UAD_BI**
  - Star – Undergraduate Admissions
UCDW Schema Contents

- **Schema Name**
- **Materialized Query Table (MQT)**
- **Tables**
- **Views**
Tables are either dimensions (_D) or facts (_F)
UCDW Views

- Two types of views
  - Materialized Query Tables (MQTs)
  - Views

View are tables whose definition is based on a query. The data contained within a view is derived from one or more tables on which the materialized query table is based. MQTs (_M) are physical tables and Views (_V) are dynamic – generated upon request.
Using DB Visualizer – 4/21/2017
Sign-up sheet in commons