Introduction to UCDW Star Schemas and Data Marts

Data Infrastructure IRAP Training
3/20/2017
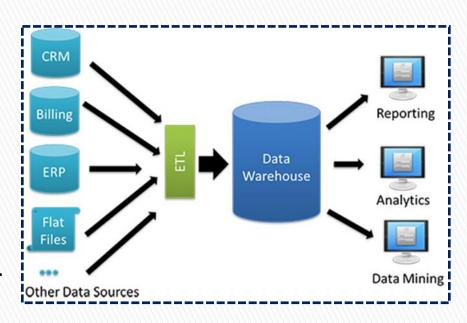
Agenda

- 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
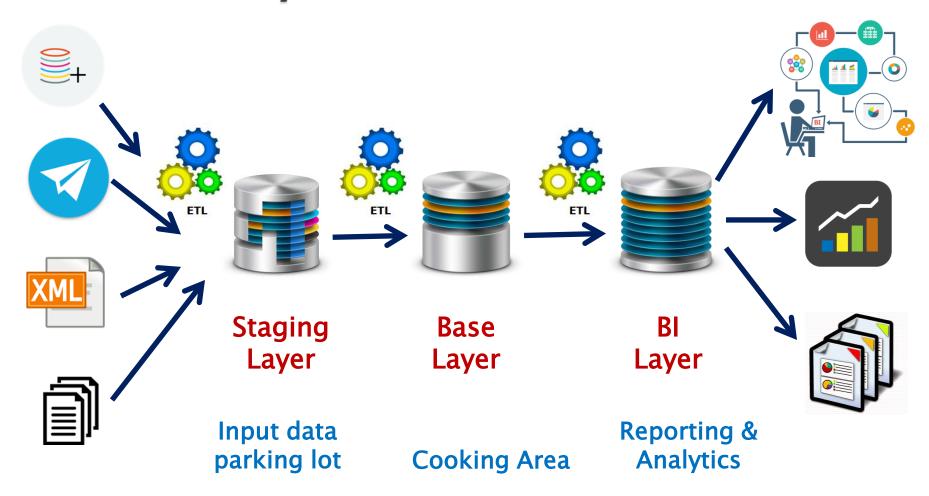
UCDW Environments
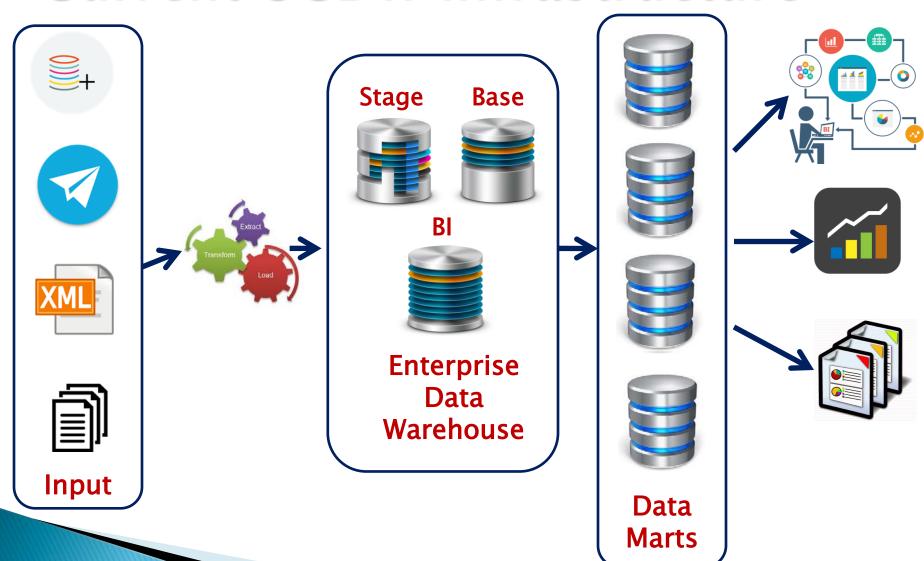

1 Development
DWD2

2 Quality Assurance
DWP3

3 Production
DWP2

Extract–Transform–Load

The 3-Layered Architecture

Staging Layer

Input data parking lot

Base Layer

Cooking Area

BI Layer

Reporting & Analytics

Current UCDW Infrastructure

Input

Stage   Base

BI

Enterprise
Data
Warehouse
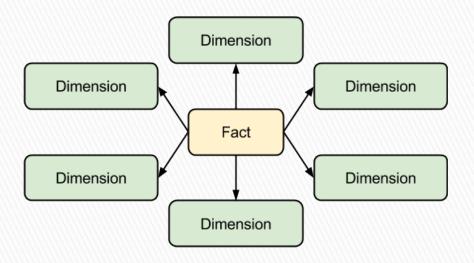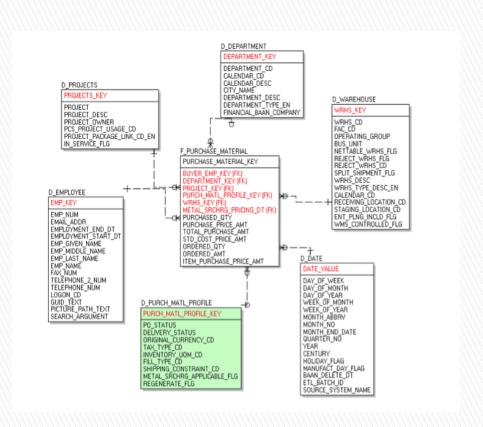
Data
Marts

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

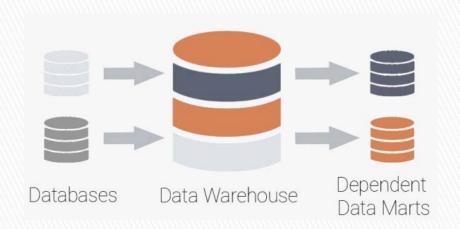▸ Data is organized into fact and dimensions

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

Databases   Data Warehouse   Dependent 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
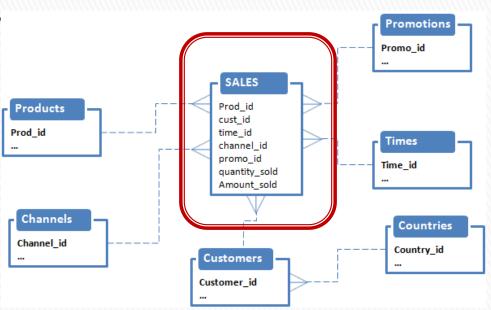- Flat table structure

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 <u>grain</u>: 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)

Dimension Table Characteristics

- 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)

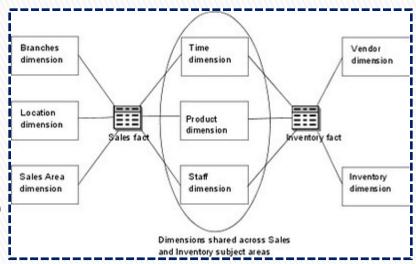# More on Facts & Dimensions

Dimensions

Time Series

Table 1a: Enrollment by Campus, Level, and Gender: General Campus and Health Sciences Combined

	Fall 2011				Fall 2012				One-year change		
	Female	Male	Unk	Total	Female	Male	Unk	Total	Female	Male	Total
San Diego	**14,320**	**15,004**	**0**	**29,324**	**13,812**	**15,247**	**0**	**29,059**	**-4%**	**2%**	**-1%**
Undergraduate	11,646	11,400	0	23,046	11,202	11,474	0	22,676	-4%	1%	-2%
Lower Division	3,595	3,059	0	6,654	3,792	3,599	0	7,391	5%	18%	11%
Upper Division	8,051	8,341	0	16,392	7,410	7,875	0	15,285	-8%	-6%	-7%
Graduate	2,319	3,228	0	5,547	2,242	3,376	0	5,618	-3%	5%	1%
1st Stage	1,897	2,593	0	4,490	1,848	2,728	0	4,576	-3%	5%	2%
2nd Stage	422	635	0	1,057	394	648	0	1,042	-7%	2%	-1%
Resident	355	376	0	731	368	397	0	765	4%	6%	5%

Facts/Measures

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

Slowly Changing Dimensions

▸ UCDW currently uses these two concepts:
  ◦ <u>SCD Type 1</u> – 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.
  ◦ <u>SCD Type 2</u> – 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.

Slowly Changing Dimensions E.g.

SCD Type 1: Student's date of birth changed from 1972-07-05 to 1975-07-05

STUD_LOC_CMP_CD	STUD_ID	STUD_CUR_ACTV_FL	STUD_BEG_EFF_DT	STUD_END_EFF_DT	STUD_DT_OF_BTH
01	15084017	Y	2002-06-30	9999-12-31	1975-07-05

SCD Type 2: Student's last name changed from Nelson to Shaw on June30th 2002

STUD_LOC_CMP_CD	STUD_ID	STUD_CUR_ACTV_FL	STUD_BEG_EFF_DT	STUD_END_EFF_DT	STUD_LST_NAM
01	15084017	N	2000-11-08	2002-06-29	NELSON

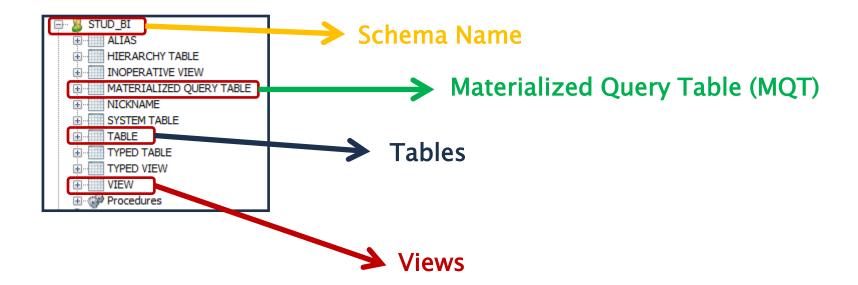| STUD_LOC_CMP_CD	STUD_ID	STUD_CUR_ACTV_FL	STUD_BEG_EFF_DT	STUD_END_EFF_DT	STUD_LST_NAM
01	15084017	Y	2002-06-30	9999-12-31	SHAW

Most recent record

UCDW Naming Conventions

- Dimension tables have a suffix of _D
  - Examples: STUDENT_D, STUDENT_LEVEL_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)
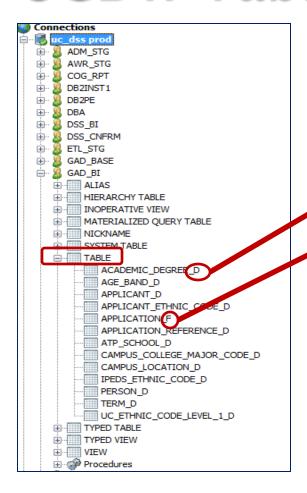 - Examples: ENROLLMENT_HEAD_COUNT_M, IPEDS_ETHNIC_CODE_V

UCDW Schemas of Interest

Connections
uc_dss prod
- ADM_STG
- AWR_STG
- COG_RPT
- DB2INST1
- DB2PE
- DBA
- **DSS_BI**
- DSS_CNFRM
- ETL_STG
- GAD_BASE
- **GAD_BI**
- GAD_STG
- GLONG_STG
- HPEJMAN
- **IRAP_BI**
- ISE_BASE
- ISE_STG
- KAORUMA
- NULLID
- SP_BASE
- SP_BI
- SP_STG
- SQLJ
- STUD_BASE
- **STUD_BI**
- STUD_STG

▸ **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**
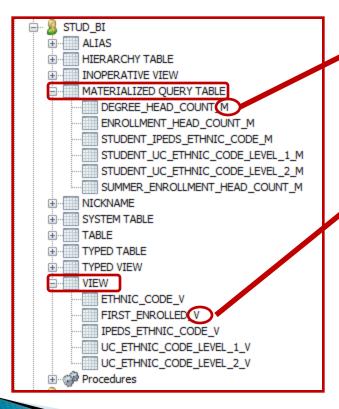 ◦ Star – Undergraduate Admissions

UCDW Schema Contents

```
☐ 🔊 STUD_BI
   ☐ ▦ ALIAS
   ☐ ▦ HIERARCHY TABLE
   ☐ ▦ INOPERATIVE VIEW
   ☐ ▦ MATERIALIZED QUERY TABLE
   ☐ ▦ NICKNAME
   ☐ ▦ SYSTEM TABLE
   ☐ ▦ TABLE
   ☐ ▦ TYPED TABLE
   ☐ ▦ TYPED VIEW
   ☐ ▦ VIEW
   ☐ 🔩 Procedures
```

→ Schema Name

→ Materialized Query Table (MQT)

→ Tables

→ Views

UCDW Tables

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