

The CSU Data Lake



California State University

Patrick Perry, CIO

Current Analytics Landscape

- CSU has a common ERP system, 3 primary modules; BI modules in place for some of these
- Lots of secondary “shadow” data collections: ERSS, APDB, Facilities, etc.
- No true central data repository/data “lake” (structured/raw)
- Many campuses have developed their own local DW/BI; others have not; many platforms
- Multiple reporting tools: SAS-BI, Oracle (OBIEE), Tableau, various dashboards and static web reports
- Not fully architected to talk to each other (matching keyfields, common identifiers); some data quality issues

Data Needs: In Support of Student Success

- Graduation Initiatives will require data to perform analysis of friction points, inefficiencies, student pathways
- Data needed to set appropriate goals and targets and monitor progress, not just at end point, but all along the way (momentum points)
- Data needed to create benchmarking and a common community of practice
- External data needed: CDE, CCC, CTC, Student Clearinghouse, EDD to show levels of preparation upon entry and licensure/workforce outcomes

Conceptual Design and Services

- Create a single “data lake” for all relevant CSU system data
 - We don’t need to change its native structure, transactional location or ownership; we need to extract data from all the data siloes currently in use and put in one location for all to use
- Setup automated extractions to move data into the “lake”
 - Start with existing static datasets; no new collections
- Anonymize data; create master pseudo-id
- Identify linking key fields and map data elements

Design and Services

- From this data “lake”:
 - Create a “semantic layer” of the data (business representation of the data that helps end users access data autonomously using common terms)
 - Value-add the data by creating easier structures to query, calculated/derived fields, and adding external data
 - Feed data back to campuses in the form of standardized datasets, query capability, app development capability (central and local), reporting layer, dashboard layer

Design and Services

- Create a virtual query environment for campus access
 - Using VM Ware, allow authorized users at CO and campuses to have access to the data lake
 - Security through CSU identity management; access dictated by credential
 - Allow campuses to use their own tools
 - These can be placed into each virtual users environment
 - Allow data upload and “scratch” area
 - Foster the “culture of data/evidence” through professional development, partnerships inside CSU

Design and Services

- Use data “lake” as the back end for CO-delivered apps, dashboards, and reporting
 - Creates standard definitions of metrics
 - Allows for automatic updating of all apps when source data are modified

Design and Services

- Create a single access web portal for all CSU data
 - “data.calstate.edu”
- Can point to any referenced data in CSU, but people don't have to hunt for it
- Public data open; private data secured by credential

Early Objectives

- Start with internal pilot first, using existing historical data
- Evaluate needs, platforms, resource needs, and available resources
- Data Governance Structures:
 - Review security and privacy issues
 - Discuss issues surrounding campus access to other campus' data
 - Determine researcher access protocols
 - Identify future development efforts

Progress To-Date (Architecture)

- Deployed SQL Server 2016 as Data Platform
- Selected Tableau as the Reporting Platform
- Installed Tableau Server
- Planning Stage:
 - Security
 - Authentication & Authorization
 - Workflow
 - Content Review and Publishing

Progress To-Date - (Data Loaded)

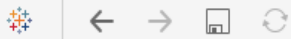
- Enrollment Reporting System files (ERSx)
 - Applicants -2007 to 2016
 - Degrees -2008 to 2016
 - Students -2008 to 2016
 - Special Session -2008 to 2016
 - Race Ethnicity -2009 to 2016
 - Teacher Credentials -2006 to 2016
- Financial Information Resource Management System files (FIRMS)
- Academic Planning Data Base (APDB)
- Spaces & Facility Data Base (SFDB)
- EDD Wage data
- Common/COSAR table
- Degrees Data Base

Progress To-Date - (Development)

- Discussions with CCC and CDE for data sharing and common CSU ID
- Adding Descriptions to ERSx and FIRMS data
- Adding Derived Fields
- Planning Stages for Teacher Preparation Dashboard
- Preliminary Student Dashboard development

Tableau - Book2

File Data Server Window Help



Connections

Add

pwbmwsqldb01.co.calstate.edu
Microsoft SQL Server

Database

Select Database

Enter database name

- COSAR
- DATA
- EDD
- ERS
- ERS_Pinnacle
- FADB
- FIRMS
- master
- msdb
- tempdb

pwbmwsqldb01.co.calstate.edu

Sort fields Data source order

Show aliases Show h

Tableau - Book1

File Data Server Window Help

← → 🖨️ ↻

Connections Add

pwbmwsqldb01.co.calstate.edu
Microsoft SQL Server

Database

DATA

Table

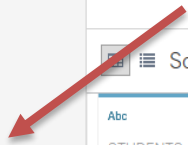
- APPLICANTS (ERS.APPLICANTS)
- DEGREES (ERS.DEGREES)
- FINANCIAL_AID (ERS.FINANCIAL_AID)
- PELL_STATUS (FADB.PELL_STATUS)
- STUDENTS (ERS.STUDENTS)
- STUDENTS_OFF_CAMPUS (ERS.STUDENTS_OFF_CAMPUS)
- STUDENTS_PROFILE (ERS.STUDENTS_PROFILE)
- STUDENTS_PROFILE_TRIMMED (ERS.STUDENTS_PROFILE_TRIMMED)**
- STUDENTS_RACE_ETHNICITY (ERS.STUDENTS_RACE_ETHNICITY)
- STUDENTS_SPECIAL_SESSION (ERS.STUDENTS_SPECIAL_SESSION)
- TEACHERS_CREDENTIALS (ERS.TEACHERS_CREDENTIALS)
- WAGES (EDD.WAGES)
- WAGES_RANGED (EDD.WAGES_RANGED)
- New Custom SQL

STUDENTS_PROFILE_TRIMMED (ERS.STUDENTS_PROFILE_TRIMMED)

STUDENTS_PROFILE_TRIMMED

Sort fields Data source order Show aliases

Erssp Year	Erssp Term	Erssp Term N...	Erssp Campus	Erssp Campus...	Erssp Birth Da...	Erssp	
2007	4	Fall	20	Chico	19700725	M	
2007	4	Fall	20	Chico	19711202	M	
2007	4	Fall	Fall	20	Chico	19711229	M
2007	4	Fall	20	Chico	19760325	M	
2007	4	Fall	20	Chico	19791208	M	
2007	4	Fall	20	Chico	19821216	M	
2007	4	Fall	20	Chico	19830506	M	
2007	4	Fall	20	Chico	19830713	M	
2007	4	Fall	20	Chico	19840223	M	
2007	4	Fall	20	Chico	19840623	M	



MAIN | HeadCount | Age | Ethnicity | Origin



Enrollment Home ▶

Select Reporting values

HeadCount

Select Campus

(All)

Year of Select Year

(All)

Select Admission Basis

(All)

Select Student Type

(All)

Select Residence Status

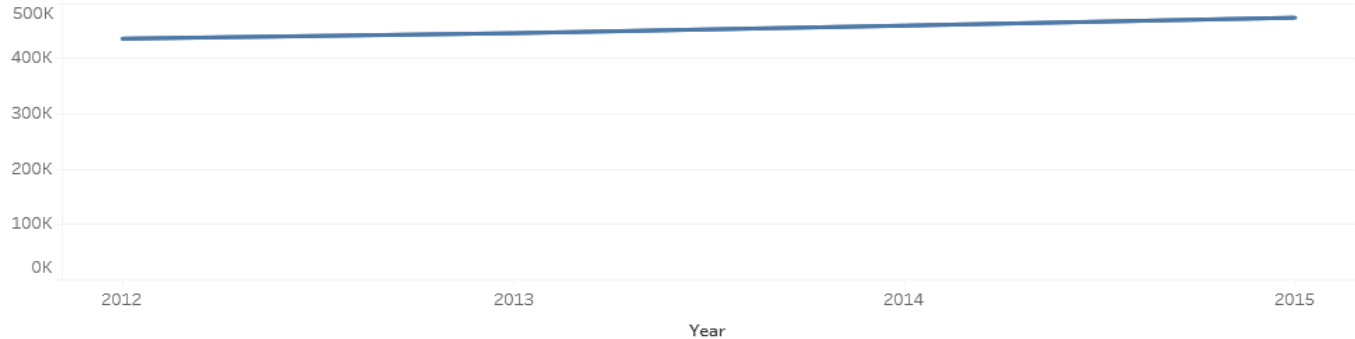
(All)

Select Student Level

(All)

CSU Total HeadCount by Campus, Fall Profile

3/28/2017 3:54:08 PM



CSU Total HeadCount by Campus, Fall Profile

3/28/2017 3:54:08 PM

	2012	2013	2014	2015
Grand Total	436,560	446,530	460,200	474,572
Bakersfield	8,520	8,371	8,720	9,228
CalStateTEACH	569	638	794	846
Channel Islands	4,920	5,140	5,879	6,167
Chico	16,470	16,356	17,287	17,220
Dominguez Hills	13,933	14,670	14,687	14,635
East Bay	13,851	14,526	14,823	15,528
Fresno	22,565	23,060	23,179	24,136
Fullerton	37,677	38,325	38,128	38,948
Humboldt	8,116	8,293	8,485	8,790
Intl Programs	493	527	509	487
Long Beach	36,279	35,586	36,809	37,446
Los Angeles	21,755	23,258	24,488	27,681
Maritime Academy	973	1,046	1,047	1,075
Monterey Bay	5,609	5,732	6,631	7,102
Northridge	36,164	38,310	40,131	41,548
Pomona	22,156	22,501	23,966	23,717
Sacramento	28,539	28,811	29,349	30,284
San Bernardino	18,234	18,398	18,952	20,024
San Diego	31,597	32,759	33,483	34,254
San Francisco	30,500	29,905	29,465	30,256

MAIN | HeadCount | Age | Ethnicity | Origin



[Enrollment Home](#) ▶

Select Campus
(All) ▼

Select Year
(All) ▼

Select Residence Status
(All) ▼

Select Admission Basis
(All) ▼

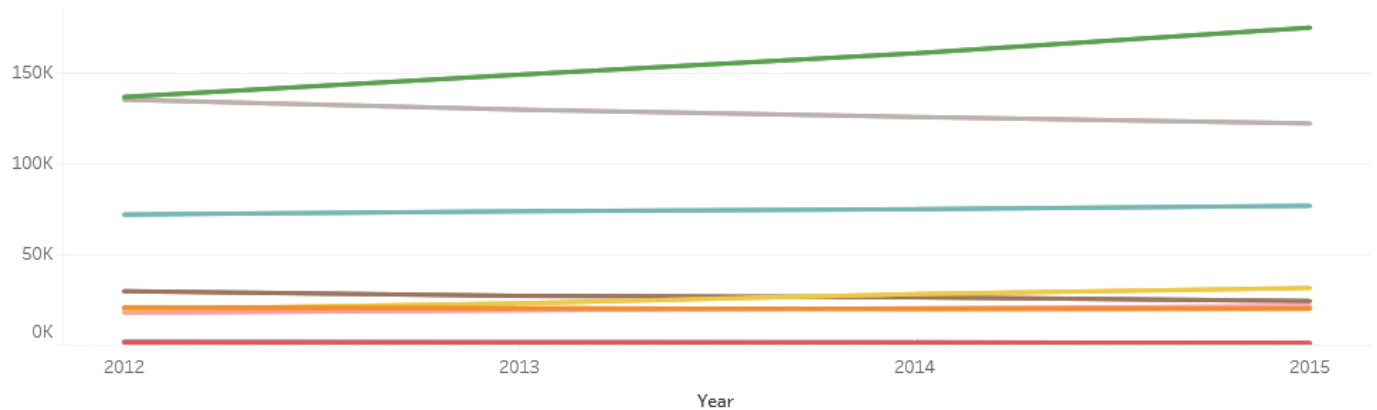
Select Student Type
(All) ▼

Select Gender
(All) ▼

Select Student Level
(All) ▼

CSU Total Enrollment By Ethnicity, Fall Profile

3/28/2017 3:54:08 PM



Ethnicity

- African Americ...
- Asian Only
- Nonresident Ali...
- Two or More R...
- White Only
- American India...
- Hispanic / Latino
- Pacific Islander...
- Unknown

CSU Total Enrollment By Ethnicity, Fall Profile

3/28/2017 3:54:08 PM

Ethnicity	2012	2013	2014	2015
Grand Total	436,560	446,530	460,200	474,572
White Only	135,335	129,838	125,835	122,273
African American Only	20,906	20,499	20,008	20,155
American Indian Only	1,642	1,481	1,420	1,201
Asian Only	72,072	73,883	75,038	76,904
Pacific Islander Only	2,012	1,850	1,757	1,299
Two or More Races	18,051	19,361	20,438	21,611
Hispanic / Latino	136,863	149,137	160,891	175,018
Unknown	29,852	27,289	26,520	24,470
Nonresident Alien	19,827	23,192	28,293	31,641

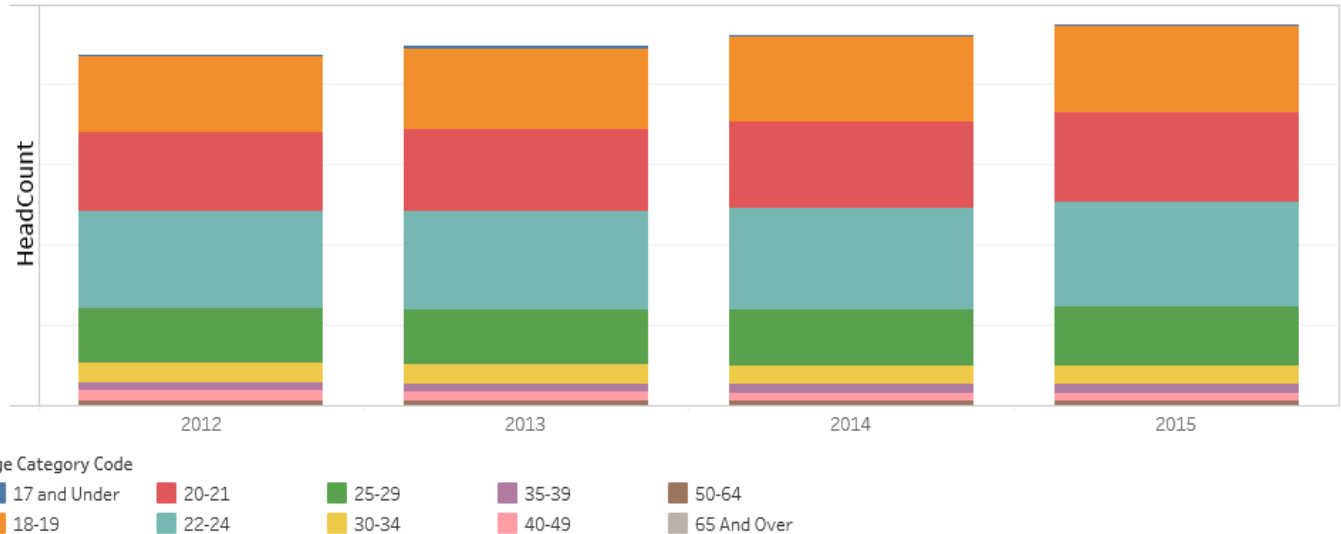
MAIN | **HeadCount** | Age | Ethnicity | Origin



[Enrollment Home](#) ▶

CSU Total Enrollment by Age, Fall Profile

3/28/2017 3:54:08 PM



Select Campus
(All)

Select Year
(All)

Select Admission Basis
(All)

Select Student Type
(All)

Select Residence Status
(All)

Select Gender
(All)

Select Student Level
(All)

CSU Total HeadCount by Age, Fall Profile

3/28/2017 3:54:08 PM

Age Category Code	Year			
	2012	2013	2014	2015
Grand Total	436,560	446,530	460,200	474,572
17 and Under	2,129	2,363	2,465	2,121
18-19	95,537	100,926	105,656	108,102
20-21	97,920	101,012	106,279	111,874
22-24	121,076	124,061	127,598	130,344
25-29	66,956	67,428	68,895	72,847
30-34	23,941	23,488	22,960	23,272
35-39	10,841	10,544	10,375	10,605
40-49	11,598	10,806	10,380	10,066
50-64	6,081	5,437	5,154	4,928
65 And Over	481	465	438	413

The screenshot displays the SAS environment with the following components:

- Code Editor (Untitled - Notepad):** Contains SAS code for connecting to an ODBC data source and querying a table of views.


```
Proc sql;
connect to odbc (dsn='Data');
create table DataTables as select * from connection to odbc(ODBC::SQLTables)
quit ;

LIBNAME SQL ODBC NOPROMPT="DSN=Data";
Proc sql;
  select * from DataTables where TABLE_SCHEM<>'sys' AND TABLE_TYPE='VIEW'
```
- ODBC Data Source Administrator (64-bit):** Shows the 'User Data Sources' tab with a single entry:

Name	Platform	Driver
Data	32/64-bit	SQL Server Native Client 10.0
- SAS - [Results Viewer - SAS Output]:** Shows the execution results in a table titled 'The SAS System'.

TABLE_CAT	TABLE_SCHEM	TABLE_NAME	TABLE_TYPE	REMARKS
DATA	EDD	WAGES	VIEW	
DATA	EDD	WAGES_RANGED	VIEW	
DATA	ERS	APPLICANTS	VIEW	
DATA	ERS	DEGREES	VIEW	
DATA	ERS	FINANCIAL_AID	VIEW	
DATA	ERS	STUDENTS	VIEW	
DATA	ERS	STUDENTS_OFF_CAMPUS	VIEW	
DATA	ERS	STUDENTS_PROFILE	VIEW	
DATA	ERS	STUDENTS_PROFILE_TRIMMED	VIEW	
DATA	ERS	STUDENTS_RACE_ETHNICITY	VIEW	
DATA	ERS	STUDENTS_SPECIAL_SESSION	VIEW	
DATA	ERS	TEACHERS_CREDENTIALS	VIEW	
DATA	FADB	PELL_STATUS	VIEW	
DATA	FIRMS	ACTIVITY_PERIOD	VIEW	
DATA	FIRMS	AGENCY	VIEW	
DATA	FIRMS	APPROPRIATION	VIEW	
DATA	FIRMS	CAMPUS	VIEW	

The screenshot shows the IBM SPSS Statistics Viewer interface. On the left, the Output window displays a SQL query for a new file. The query is as follows:

```

NEW FILE.
DATASET NAME DataSet1
GET DATA
  /TYPE=ODBC
  /CONNECT='DSN=Data;De
    'Statistics Common;
  /SQL='SELECT erss_yea
    'erss_ethnic_old, e
    'erss_inst_orig, es
    'erss_stud_lev, ers
    'erss_stud_stand, e
    'erss_total_gpa, es
    'erss_dss, erss_dss
    'erss_ge_oc_state,
    'erss_es_math, erss
    'erss_cp_soc_sci, e
    'erss_act_composite
    'erss_ept_read, ers
    'erss_sat_composite
    'erss_cp_arts, erss
    'erss_spec_prog, es
    'erss_cum_ua_pc, es
    'erss_multi_race_ca
    'erss_eap_eng_stat,
    'erss_mil_dep_stat,
    'erss_rd_sat_math_s
    'erss_cum_ue_campus
    'erss_adj_stu_level
    'DATA.ERS.STUDENTS
  /ASSUMEDSTRWIDTH=255.
  
```

On the right, the Database Wizard dialog box is open, titled "Select Data". It contains the following text:

Select the fields you want to retrieve. Then click the arrow button or drag the fields to the Retrieve Fields list.

Tip: Selecting a table selects all of its fields.

The dialog box has two main sections:

- Available Tables:** A list of tables including DATA.ERS.APPLICANTS, DATA.ERS.DEGREES, DATA.ERS.FINANCIAL_AID, DATA.ERS.STUDENTS, DATA.ERS.STUDENTS_OFF_CAM, DATA.ERS.STUDENTS_PROFILE, and DATA.ERS.STUDENTS_PROFILE_... The table DATA.ERS.STUDENTS_PROFILE_... is expanded to show its fields: erssp_year, erssp_term, erssp_term_name, erssp_campus, erssp_campus_name, erssp_birth_date, erssp_sex, and erssp_ethnic_label.
- Retrieve Fields in this Order:** A list of fields from the selected table, currently showing DATA.ERS.STUDENTS_PROFILE_TRIMM (repeated three times).

At the bottom of the dialog box, there are checkboxes for "Sort field names" (unchecked), "Show:" (with options for Tables, Views, Synonyms, System tables), and navigation buttons: < Back, Next >, Finish, Cancel, and Help.

Tableau “Ad Hoc” Demo

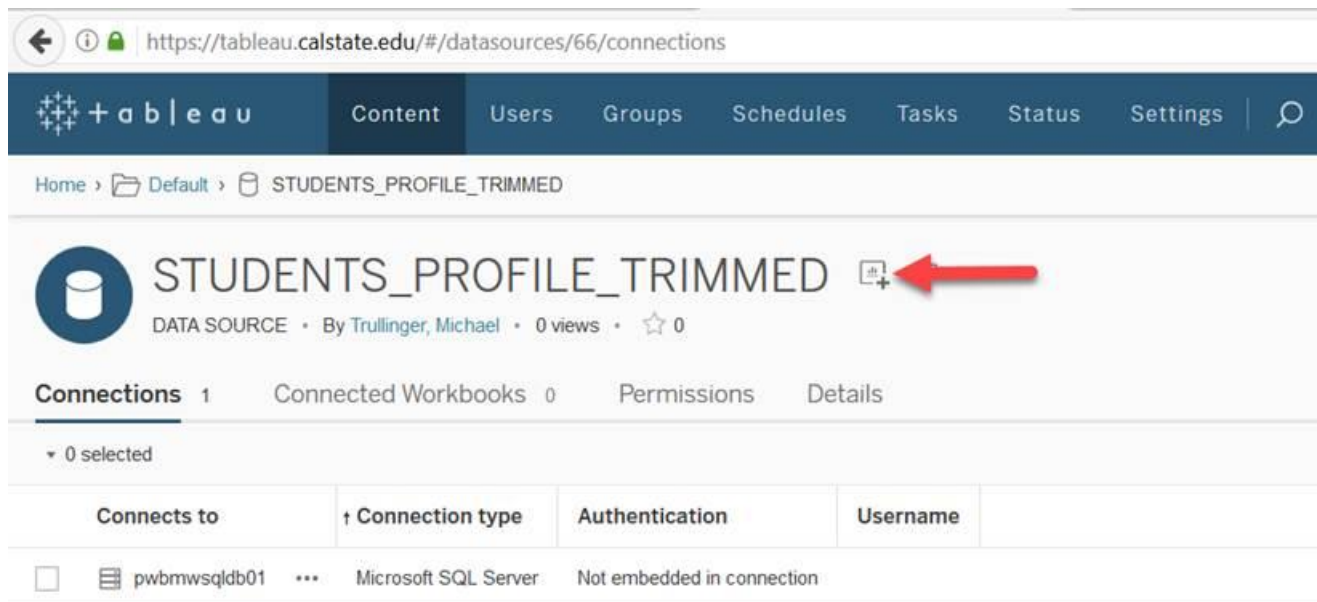


To access Tableau Web Server ad hoc workbook/worksheet developer:

#1. Login to Tableau using this link:

<https://tableau.calstate.edu/#/datasources/86/connections>

#2. Click on the 'New Workbook' button.



The screenshot shows the Tableau Web Server interface. The browser address bar displays the URL: <https://tableau.calstate.edu/#/datasources/66/connections>. The Tableau navigation bar includes 'Content', 'Users', 'Groups', 'Schedules', 'Tasks', 'Status', and 'Settings'. The breadcrumb trail is 'Home > Default > STUDENTS_PROFILE_TRIMMED'. The main content area shows the data source 'STUDENTS_PROFILE_TRIMMED' with a 'DATA SOURCE' icon, 'By Trullinger, Michael', '0 views', and '0 stars'. A red arrow points to the 'New Workbook' button (a square icon with a plus sign) next to the data source name. Below this, there are tabs for 'Connections 1', 'Connected Workbooks 0', 'Permissions', and 'Details'. The 'Connections' tab is active, showing '0 selected' items. A table lists the connection details:

Connects to	↑ Connection type	Authentication	Username
<input type="checkbox"/> pwbmwsqldb01	...	Microsoft SQL Server	Not embedded in connection

CSU

The California State University

Business Intelligence / Data Warehouse