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
The CSU Data Lake

pwbmwsqldb01.co.calstate.edu

Connections
Add

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
FAD
FIRMS
master
msdb
tempdb

Sort fields  Data source order  Show aliases  Show h
## The CSU Data Lake

### Tableau - Book 1

#### Connections
- pwbmwsqldb01.co.calstate.edu
  - Microsoft SQL Server

#### Database
- DATA

#### Table
- APPLICANTS (ERS.APPLICANTS)
- DEGREES (ERS.DEGREES)
- FINANCIAL_AID (ERS.FINANCIAL_AID)
- PELL_STATUS (FADD.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)

<table>
<thead>
<tr>
<th>Erssp Year</th>
<th>Erssp Term</th>
<th>Erssp Term Nbr</th>
<th>Erssp Campus</th>
<th>Erssp Birth Da...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
<td>Fall</td>
<td>20</td>
<td>Chico</td>
</tr>
</tbody>
</table>
### The CSU Data Lake - SAS

#### Proc SQL

```sql
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_SCHEMA<>'sys' AND TABLE_TYPE='VIEW'
```

#### Table: The SAS System

<table>
<thead>
<tr>
<th>TABLE_CAT</th>
<th>TABLE_SCHEMA</th>
<th>TABLE_NAME</th>
<th>TABLE_TYPE</th>
<th>REMARKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>EDD</td>
<td>WAGES</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>EDD</td>
<td>WAGES_RANGED</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>DEGREES</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>APPLICANTS</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>FINANCIAL_AID</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>STUDENTS</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>STUDENTS_OFF_CAMPUS</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>STUDENTS_PROFILE</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>STUDENTS_PROFILE_TRIMMED</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>STUDENTS_RACE_E Th NICITY</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>STUDENTS_SPECIAL_SESSION</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>ERS</td>
<td>TEACHERS_CREDENTIALS</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>FADB</td>
<td>PELL_STATUS</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>FRMS</td>
<td>ACTIVITY_PERIOD</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>FRMS</td>
<td>AGENCY</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>FRMS</td>
<td>APPROPRIATION</td>
<td>VIEW</td>
<td></td>
</tr>
<tr>
<td>DATA</td>
<td>FRMS</td>
<td>CAMPUS</td>
<td>VIEW</td>
<td></td>
</tr>
</tbody>
</table>
The CSU Data Lake – SPSS

Select Data

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.

Available Tables:

- DATA.ERS.APPLICANTS
- DATA.ERS.DEGREES
- DATA.ERS.FINANCIAL_AID
- DATA.ERS.STUDENTS
- DATA.ERS.STUDENTS_OFF_CAMPUS
- DATA.ERS.STUDENTS_PROFILE
- DATA.ERS.STUDENTS_PROFILE_TRIM

Retrieve Fields in this Order:

- DATA.ERS.STUDENTS_PROFILE_TRIM
- DATA.ERS.STUDENTS_PROFILE_TRIM
- DATA.ERS.STUDENTS_PROFILE_TRIM
- DATA.ERS.STUDENTS_PROFILE_TRIM

Soft field names

Show:

- Tables
- Views
- Synonyms
- System tables

April 2017 IR Meeting
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.