Introducing CCC Data
Introduction

Mark Cohen
Product Manager,
CCC Data Services & Transcripts
CCC Technology Center

mcohen@ccctechcenter.org
Agenda

- Project History
- Overview
- Data Sources
- Flow of Data
- Security
- Stakeholder Engagement
- Next Steps
- Discussion
Introducing CCC Data

The Data Warehouse is part of the CCC Data project, supported by the Data Services Program (DSP) initiative from the California Community Colleges Chancellor's Office.

- Project overseen by CCCCO Data Governance Council
  - Manages MOU’s and data sharing agreements that enable data to be stored in the Data Lake and accessed through the Data Warehouse.
• FY 2016/2017 - CCC Tech Center was asked through the Tech Center Grant to construct a Data Lake to support Canvas data as part of the Online Education Initiative (OEI), assessment data as part of the Common Assessment Initiative (CAI) and CCCApply data. An initial project to do a Proof of Concept (POC) was constructed to create the data lake and store data into Amazon's S3 object data store. That POC completed in May of 2017.

• 2017/18: After the successful completion of the POC, the OEI Workplan and the Common Assessment Grant called for the Tech Center to build a Data Warehouse to house CCCApply, MyPath, Canvas and Multiple Measures data. That work took where the Data Lake POC ended and began work in June of 2017 and the current Data Warehouse project began. The intent was to build out a production ready data warehouse and begin piloting it out with Colleges in the Spring prior to going into Production.

• FY 2018/19: the CCC Technology Center began working on an architecture for system-wide data, information, and technology infrastructure which includes Business Intelligence. We brought in CCCApply, Canvas and MyPath and de-escalated the CCCAssess data as per the instructions from the Chancellor's Office. In late May 2018 we narrowed our focus to providing all of the Colleges the LGBTQ report, based on the legislatively driven need for the LGBTQ reports, and for the need for flexibility so we can pivot to follow the greater strategy of the Chancellor's Office for the next few years. We piloted the data warehouse to build on the POC. An infrastructure was built out to
move raw data from the sources to the S3 data lake and from there into readable tables and objects. This data was then presented through the CCC Report Center interface

• FY 2019/20: completed successful pilot of the Data Warehouse, including data from CCC Apply, MyPath, and the colleges Canvas data, with Foothills, Butte, Shasta, Lake Tahoe, Yuba, Mt. SAC. We added COCI and C-ID data to the Data Lake. Now working to add MIS and Cal-PASS+ data to the Data Lake. Launching CCC Data to production, including the Data Lake, Data Warehouse, and Data Pipelines; and launching the DW Report Center to all CCC’s. Forming the Data Warehouse Advisory Group and continuing to work closely with the CCCCO and data governance.
• The CCC Data Warehouse project provides a set of data products to serve the California Community Colleges and the California Community Colleges Chancellor's office, this includes:
• A Data Lake in which any of the source data can be persisted as it comes from the source for data mining and auditing purposes. All data from source database are stored with changes of data over time.
• A Data Warehouse that acts as a structured source of master data that can be used to generate the data marts, reports, and analytics that the end-users need. Holds unencrypted data available for researchers.
  ○ 115+1 schemas ensures security, creating distinct schemas against the data warehouse so that each college accesses only their data.
• A report center, and ODBC/JDBC connections directly to Redshift that provide colleges and Chancellor's Office with access to these data for reporting and analytics.
• The ETL’s, data pipelines, that support the movement of data between the data sources and CCC Data tools.
• Overtime, this may expand to include additional elements, we will discuss later in the presentation
### Current Data Sources

<table>
<thead>
<tr>
<th>Source Data</th>
<th>Data Lake</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCC Apply: Application</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CCC Apply: International Application</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CCC Apply: Fee Waiver</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Multiple Measures Placement (MMPP)</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>MyPath</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>COCI</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>C-ID</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>MIS</td>
<td>✔</td>
<td>Pending Approval</td>
</tr>
<tr>
<td>Cal-PASS+</td>
<td>Pending MOU</td>
<td>Pending Approval</td>
</tr>
<tr>
<td>Canvas (integration per college)</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

- Connected to the DL through a series of ETL's
• CCC Data is built on an AWS-centric, cloud-based solution that stores and structures data sets from the CCC data sources.
• Source data consisting of system data, and college specific data
• Connected through a series of ETL’s that are developed based on requirements
• The data pipelines run nightly, or more often based on requirements
• All data is captured in the datalake
• Value is added through capturing incremental updates to the data, identifying which which data sources have deltas/diff info generated so Researchers can find data, that in some cases, only exists in the DW as its over written in the source db
  o DIFF tables captured for for CCCApply (standard, Fee waiver and international), C-ID and COCI
  o Do not store diff tables for MMPS, Mypath & Canvas at this time due to use case or nature of how data is shared
• The data is quickly discovered, retrieved, and used for analytics, reporting, or data mining.
• Data Warehouse enables colleges to perform analysis across their data and system data
  o AWS Redshift supports up to 1,024 unique schemas against the data repository, effectively creating 115+1 distinct data warehouses, assuring that each college accesses only their data
In addition to the Report Center as the default method to get to the data, researchers can connect to Redshift via ODBC/JDBC to use their own tools. (Power BI, Cognos, etc.)
- AWS Data Pipelines are used for pulling data from the source database.
- Using AWS data pipeline to talk to external data sources and pull it into the DL, from there data pipeline brings data from DL to DW.
- AWS data pipelines developed and maintained by CCC Data Team
- We control when it runs and with what frequency, source DB doesn't know when shared and coming into DL.
- SuperGlue dumps data from source database into the Data Lake
- SuperGLue team, responsible for moving data from the data source to the API gateway...
- SuperGLue controls timing, knows what database, how often, how to fetch and bring to DL.
- Streaming data arrives at Data Lake via AWS Kinesis
- MyPath uses Kinesis, developed as part of MyPath architecture.
- Source team gets log data, publishes the data to Kinesis, and Kinesis moves the data to the DL.
CCC Data Built on AWS Platform

- AWS selected through an open RFP process
- Flexible architecture is highly scalable
- Relatively low cost and transparent pricing
- Address speed, performance, and storage requirements
- Includes robust security and supporting tools
- Leverages existing AWS infrastructure
- 24/7 Monitoring and Incident Response
Amazon Web Services provides one of the most secure cloud environments available for sensitive data and confirms to the following standards:
Securing CCC Data

- Individual instances for each College (horizontal segregation)
  - Colleges may only access the data associated with their institution
- PII
  - All data encrypted at rest
- Adhere to Standards:
  - NIST guidelines
  - AWS best practices
  - FERPA and CA regulations
- All transactions logged and monitored
- Report Center Integration with Infrastructure:
  - Accessed through SSO
  - Connected through (IDP) local Identity Management (IDP), SAML-based
Stakeholder Engagement

Data Warehouse Advisory Group

- Advisory group made up of
  - CCC institutional research, planning and effectiveness Professionals (identified by the RP Group)
  - Chancellor's Office stakeholders
- Meet monthly to inform and help prioritize requirements:
  - Identify system and college data sources
  - Requirements for Report Center reports
  - Colleges connection to their data

- The goal for this advisory group is to
  - inform the business requirements for how colleges connect to the Data Warehouse, including the data accessed that support reporting, analysis, and research, and to
  - provide guidance to ensure that this project is developed in a manner consistent with the needs of the CCC IRPE community.
- CCC Data project guided by their input, requirements.
- This Advisory Group is composed of institutional research, planning and effectiveness (IRPE) professionals from California Community Colleges, along with representation from the Chancellor's Office and CCC Technology Center.
Next Steps

• Engage Data Warehouse Advisory Group
• Continue coordination with CCCCO efforts
• Connect more data sources
• Support colleges accessing Data Warehouse & Report Center
• Ongoing development of CCC Data infrastructure
• Develop reporting in Data Warehouse Report Center
• Participate in Data Services Program activities

• Engage Data Warehouse Advisory Group
  identify requirements, data sources, reporting for Data
  Warehouse, and DW Report Center
• Work with CO on continued direction of CCC Data, MOU’s governing
  data sources, data policies, and data related use cases
• Continue developing data pipelines to bring more data in to the DL, DW,
  Report Center
• Work with colleges connecting to the DW through DW Report Center
  and through ODBC/JDBC to Redshift,
• Develop reports in the DW Report Center working with advisory group
  and other stakeholders
• Work with CO on systemwide data projects
Based on direction from the CCCCO and input from the advisory group, future development of CCC Data, may include:

- Bring in more sources of college data, so that the DL and DW are made up of both data originating from the colleges as well as system data.
- Develop a set of Data Marts from the Data Warehouse and Data Lake, to provide zone- and domain-scoped data sets with dashboards, reports, and analytics targeted to those users.
- Evaluate business intelligence tools that may expand on the functionality of the CCC Data platform.
- May explore multi-dimensional/Cube (OLAP) databases as needed.
Related DSP Activities

In coordination with the Chancellor's Office Digital Innovation and Infrastructure Division:

- Participate in CCC Data Governance Council
- Support selection of systemwide Data Dictionary Application
- Develop strategy for Master Data Management
Data Warehouse Access

- CCC Data available to CCC Institutional Research, Planning and Effectiveness
- Data Warehouse Report Center
  - Upgraded Report Center will be available to Researchers with access to LGBTQ report
  - Addtl staff may request access at staffsupportccctc@openccc.zendesk.com
- ODBC/JDBC Connection
  - Colleges may request access to the CCC Data Warehouse through request to staffsupportccctc@openccc.zendesk.com
Discussion
Introducing the Data Warehouse

Mark Cohen
mcohen@ccctechcenter.org