CCCCO System-wide Architecture Committee (SAC)
Meeting Minutes
Friday December 15, 2017
Zoom Online Meeting

Attendees: Lou Delzompo, Joseph Moreau, Jeff Holden, Ben Seaberry, Sylvia Lynch, Robert Hughes, Dave Fuhrmann, Alex Jackl, and Bruce Racheter

Call to Order:

Lou Delzompo called the meeting to order at 1:30 pm and took attendance.

Minutes:

The meeting minutes for October 26, 2017 were approved with no changes or corrections.

Data Lake / Data Warehouse:

Alex Jackl provided an overview of development and the structure of the data lake and data warehouse that went into release today. They will be opening it to researchers once the post review security analysis is done and there are a couple of conversations with the Chancellor's Office and leadership.

Security is paramount. Therefore, separate schemas are being built for separate colleges with horizontal segregation. This keeps data in different schema so colleges are not accessing the same segment. The project is also integrating work with the core administrator project to there can be single-sign-on (SSO) and identity management to allow identification of particular roles. Some roles will be permitted access to PII while others will not. Roles will be driven by the colleges.

The data warehouse will include data from each of the four prototype Canvas instances, MyPath data, and CCCApply data. The project is also constructing a minimum data dictionary so they can attach metadata, for example whether data is sensitive or not, when it was last updated, etc. This will allow them to maintain their data, understanding what each of the columns in the technical data structures mean. That will be rolled out over the next couple of months.

The data warehouse is a combination of five things including a series of connectors between various data sources into the data lake. It is an S3 data structure designed to be where all the data is placed and kept as it is being collected. It is for storage and is almost archival. People will not be able to access the data lake directly. There will be an ETL that moves things from the data lake into RedShift data tables that are columnar structures. The data tables are basically columns and rows in a dimensional format. So there can be dimension tables off of fact tables. Those are currently connected to the Jaspersoft reporting software that runs in the CCC Report Center, which is the same kind of reporting structure used for CCCApply.

The Data Warehouse is being designed from the start to be easily integrated. They are using Glue components now. Later, when SuperGlue with master data management, and various connectors like the College Adaptor have been built inside the Glue framework, they will natively be able to connect up to the Data Warehouse.

For the foreseeable future, the Data Lake is not intended to be accessible to anyone except Technology Center or Operations staff. It is meant to be the collection point of a lot of information that may at some point be useful. Perhaps for an audit of the data, or being able to prove to a student that no one changed their grade, etc. The Data Lake will not be accessible to just anyone, and they will only make accessible in the data mart the things that are pertinent to the Chancellor's Office, researchers, etc., who are given access.

The current thought is that a lot of fast interconnectedness would be made accessible through SuperGlue directly from the data sources, and the Data Warehouse would store vetted long term data. SAC will probably have more conversations in the future about how to best connect applications together. There are two different streams, applications talking to each other and also applications putting data into long term storage in the data warehouse.

The CCCApply data source is using an AWS pipeline. This is basically a choreography engine that allows execution of scripts to move data from various different types of sources into various different types of sources. In this case there is an AWS data pipeline for Apply in the CCC Technology Center owned environment of Amazon. It talks to CCCApply which is also a Tech Center application, but could be anything. The data pipeline is pulling scripts from Apply and pulling into the S3 into the data lake. At that point there is an engine that AWS supports called EMR that is part of its data mining/big data construct that runs scripts to synch it up. They check to see if it is new data, correcting data, or an add-on. There are structures that can check in the data warehouse and the data lake and do the appropriate action to update records or add new records based on the script.

With the MyPath data the team started with the event logs, "This user accessed this application at this time." They are pulling data through Kinesis Firehose sitting on top of MyPath. The advantage there is the ability to pull it in much faster than with nightly builds. Right now it is set slow, but it can be dialed up to do live pulls if desired. That data drops into AWS 3 and then goes through the same process with appropriate data going into the reporting store.

MyPath and CCCApply are centralized services, so all the college information is in one application and one data structure. However, with Canvas there are individual Canvas environments for each college, so there is a Glue connector service. Canvas has a set of APIs that are like little micro-services that pump out data from Canvas. The project has built service workers that monitor changes or at an appropriate time the service worker will pull the data into Glue which will then write that into the S3 buckets in the data lake and also transfer it over to Redshift at the appropriate time and in the appropriate quantity.

The onboarding process for students might change with MyPath. A student profile service can used to orchestrate some work flows in the student onboarding process. That student profile service inside of MyPath may also benefit from the Kinesis mechanism where it will keep the student profile service up to date at all times.

In the AWS Redshift tables the last thing that happens in Jaspersoft is that data domains are created for the appropriate end users. There is a MyPath data domain and five data domains for CCCApply (including the International application data, and the College Promise domain).

There are two levels of filter that can be used to control what data are put into Redshift to move into the long term data store, what domains get created in Jasper, and what is visible and reportable. Right now it is very primitive with just a super user and a college user. But Alex expects over the next couple of months after they build the security profile a college or district will be able to identify who can see PII and who can only see aggregated data.

The pilot release of 1.1 is targeted for March/April. The key purpose of that release is to add role based security and permissions. It is important to secure data properly and give access properly. That release will also increase the scope of the personas based on the use cases the colleges provide. There may also be a return to inclusion of multiple measures, but that is up in the air right now. Access to multiple measures data was removed because there were decisions that still needed to be made about whether the CalPASS algorithm or the algorithm built into the CCCAssess platform would be used. There was enough uncertainly around it that the team was asked not to make any of the current data available. Data structures were built out to support inclusion of multiple measures, but they are currently disabled. The project will wait on a decision from the appropriate decision makers.

Prototype colleges currently giving feedback are Foothill, Butte, Lake Tahoe, and Shasta. The plan is for 1.1 to expand the number of pilots. They will work with SAC and the Chancellor’s Office to come up with the right number to include for that release. The intention is to move toward a production release in June which will be open to many more colleges.

They will be constantly working on making the reporting interface better. Jaspersoft is not ideal for a data warehouse structure, but they are working with it for now to create reporting interfaces useable by researchers. Simultaneously they are researching other reporting and analytics tool options that allow more data warehouse capabilities like cross domain reporting and more advanced visualization; those are not in the current report set. However, the most common use case researchers will use is to create queries and then download and work with data in Tableau or whatever other tool they use locally.

The charter as defined by Joe Moreau and the Foothill team is about the research component. There may be opportunities to do other things, but at this point they haven't been assigned them. This data warehouse project started as part of CCCAssess and before it was cancelled the OEI team asked that Canvas data be captured to facilitate colleges' access. Ben suggested there is a lot of energy and interest in multiple measures and other ways for students to get into classes without needing an assessment. Lou recommended Ben pass his thoughts on to Laura Hope.

Joe suggested getting the data warehouse it into the hands of researchers so they could work with it and come up with better ideas about what would be most useful for them. Alex cautioned there is a technical issue related to security if each college has their own tool they want to use to pull data from the data warehouse. All of the security structures have to be in place with all the security protocols to identify data, discrete college data, and make sure it is protected. Building a generic protected data structure would be challenging and require configuration of the tools as well. The team is looking at whether they can have an interface that connects to whatever local tool is available. They are still exploring and experimenting, that is what release 1.1 is about.

Release 1.2 is about expanding by doing more of the same things. One of the new things that might come in is to build out the data model and data warehouse structure so it can handle e-transcript data and student profile data coming from MyPath.

There is a handshake agreement with Eagle/Aries, the company that has a lot of K-12 student data, as well work on a deal with CDE to get access to K-12 data and the SSID. This could supplement multiple measures data. E-Transcript California is also working on augmentation for new industry standards. Alex doesn't know how fast they might be asked to do the ETLs and connectors, but they are going to build out the data lake and data warehouse so they have a place to put K-12 and e-transcript data and a place to move it, since it may start showing up.

People have been talking about expanded business intelligence capabilities, but that is not currently on the work plan. The project is going to build out a data dictionary structure and some pages so it will be possible to see the elements, their definitions, and their metadata. Actually building out a data dictionary application with services so that tools can actually pull the information out of the data dictionary with restful APIs, etc. is beyond the current work plan. That would then be connected up and either stored in the data warehouse or connected up through the SuperGlue connector. Another element that has yet to be tackled is the MIS data and it is large. Researchers are definitely interested in comparing the MIS data with Canvas and MyPath data. There is an interest in that, so the team is looking at maybe doing a pilot of MIS data in FY 17/18 and integrating that into the data warehouse. This is just a sample of what people are asking about for the next fiscal year for the data warehouse.

Lou talked briefly about Data Management/Data Governance concerns. There has been some back and forth with the Chancellor's Office about the need for a group to provide oversight into decisions. That could have to do with who should have access to what kinds of data and what the approved uses of data would be. One example of a potential concerns could come from use of Canvas data to evaluate professors and/or instructors. However, right now there is not a team in place to make those decisions and the Technology Center has been consistent in saying it shouldn't be the one making them.

Lou felt the Chancellor's Office rejected the Data Management Office (DMO) and Data Management Advisory Committee (DMAC) as being too much oversight; right now the DMO is really Alex working with the four researchers. The idea was for SAC to be involved is some of this with representation in both the DMO and DMAC and TTAC would have representation as well. Lou felt decisions about whether or not to include multiple measures in the data warehouse could have been facilitated if the data management structure was in place.

Joe Moreau thought combining the data management framework particularly in combination with TTAC made sense as a good place to start. He thought a reason for resistance to setting up data management was lack of understanding of what the structure would recommend or prohibit. He suggested developing use cases or case studies to demonstrate the kinds of issues or decisions data management could help resolve. "This is how the researchers would like to use it, and these are some decisions that need to be made about those uses." Those use cases might help people understand and then they could provide feedback.

The team has tried to be educated by what other people have done. They have tried to listen and come up with possible approaches. It is challenging because the data being collected for MyPath is largely unknown to anyone. Similarly, Canvas data is largely unpublished and unknown. The Chancellor's Office is used to MIS and CCCApply; they know what is in those data sets. For example, they are excited about access to the LGBTQ data. This will allow college researchers better access to that data.

Members discussed possible causes of paranoia about data and uses of data and whether or not that might be an element in resistance.

Another benefit of use cases is to be able to show how, "One college found great benefit from using this data in this way." Those success stories are possible when researchers are able to bring together information for better quality interventions or student success or retention, etc. Showing the benefit far outweighs the risk is important in making the Chancellor's Office feel more comfortable and see tangible benefits. They are concerned and reluctant, so give them good examples they can learn from. To the degree that they can understand it and see it, that will help them support a data management framework.

Joe agreed, and thought it would be helpful to have those ideas come from the field rather than from the Technology Center or the Chancellor's Office. Colleges want to know what other colleges are doing and how they are succeeding. Ideas that flow from the field have different status.

System Updates:

Security Center Update

The Spirion Software contract has been passed and approved. The purchase order was created, paid for, and now they have it. That will be available the start of the year. Jeff can make that available sooner to members who are interested. Ben is interested in Spirion.

Splunk, Tenable Security Center, and free SSL Certificates services are all available as well.

Jeff is making a list of colleges interested in security assessments. They are doing a CIS gap analysis for schools.

The Winter Information Security and Accessibility Workshop will be held January 8th and 9th at San Jose Evergreen. Anyone staying for both days can get reimbursement of up to $200 for their hotel.

Technology Center Update:

A lot of releases went out both yesterday and today. The Data Warehouse was pushed to a production instance although no one has access. The new version of CCCApply went into pilot today. Colleges are being encouraged to check that out in advance of the production release January 12th. There is new nomenclature for College Promise (in place of BOG fee waiver). They are also pushing a patch to address some of the fake applications that have been submitted. There is a machine learning project going on to build a kind of firewall to intercept what are believed to be fake applications and quarantine them so colleges don't have as much overhead. Some colleges are getting quite a few. They aren't completely sure of the reason for the fake applications, but the colleges being attacked seem to be ones that give out an email address with .edu. There may be an IRS education tax credit that can be accessed with a .edu email address. The Technology Center discovered an inordinately large number of fake applications are filed in under ninety seconds, which would eliminate a human being as the source of the application. The patch going out will automatically quarantine anything submitted in less than ninety seconds. Later there may be more advanced algorithms to look for combinations of social security numbers that have shown up on the dark web or names seen before, etc. Colleges will then be able to go in and decide if there are any that were quarantined they want to take.

A new version of CO-CI was also moved out that addresses a large number of current issues practitioners have complained about. A set of data reports were run that show people were able to get their work done, but there were still some unhappy practitioners, so hopefully this will address some of those issues.

Last week colleges had a successful User Acceptance Training of Course Exchange version 2.0. They are proposing that go into production in January.

The Technology Center is heavily involved in changes in MyPath to facilitate the first pillar of Guided Pathways; the onboarding piece. Omid has been visionary in combining pieces into a more coherent workflow. He noticed that of about 2.1 million applications submitted to the CCC, only about 700,000 students end up in classes. He is really interested in looking at what can be done to keep from losing those students. MyPath is an important piece of that. The team is also looking at mobile solutions, with an app for MyPath instead of just a page that is mobile responsive. One option is the CCC Foundation which has a mobile group, another is DubLabs.

Omid has authorized a project taking the California Promise data (formerly the BOG fee waiver) and building on what has been built in the Course Exchange 2.0 in order to write that data directly into the SIS as the student finished the BOG application. This would in some ways replace the download client. The team is starting a prototype. In Course Exchange 2.0 if a home college already has the BOG fee waiver, it passes the data over to the teaching college. With Course Exchange 3.0, will be the ability to automatically pass the CCCApply application data. That work is under development. The goal is to bypass the download client and provide more of a real time feel to those applications. Every college handles it a different way so it will be a technical challenge to undertake this work.

Next Meeting:

The next meeting is scheduled for Thursday January 25, 2018 at 1:30 pm.

Adjournment:

The meeting was adjourned at 3:00 pm.