Technology & Security Subcommittee Meeting Summary

November 19, 2020

This document provides a summary of key points that emerged over the course of the meeting. More information about the meeting, including materials, the PowerPoint, and a meeting recording are available at https://cadatasystem.wested.org/meeting-information/common-identifier-subcommittee.

The November 2020 meeting had the following goals:

- Provide an update on key decisions by the Cradle-to-Career Workgroup
- Review estimates for the number of data system records
- Learn about the UC San Diego Super Computer Center's secure data enclave
- Identify key technology components and pricing considerations for the data system

The following representatives attended the meeting:

Formeka Dent, Antelope Valley Union High School District; Helen Norris, Association of Independent California Colleges and Universities; Barney Gomez and Daryl Lal, California Community College Chancellor's Office; Alan Nakahara and Rodney Okamoto, California Department of Education; Karissa Vidamo, California Department of Social Services; Janet Buehler and Michele Robinson, California Department of Technology; Lloyd Indig, California Health and Human Services Agency; Amy Fong and Greg Scull, California School Information Services; Subash D'Souza, California State University Chancellor's Office; Matthew Linzer & Hooman Pejman, University of California Office of the President

Update on Key Decisions by the Cradle-to-Career Workgroup

The meeting opened Kathy Booth of WestEd providing an update on decisions made by the workgroup at October meeting, including approving the data request process, the core public content of the data system, and the responsibilities of the managing entity.

Then, subcommittee members were asked to volunteer for two homework teams to develop:

- **A permission protocol**: will be used to allow each partner entity to designate who within their organization has access to the information in the cloud and, once developed, the secure research environment. The managing entity will also have access to this information to support state data system activities.
- **A data deidentification process**: How data is deidentified for public release—determining things like statistical masking, procedures for handling the deidentification of data, how much data will be masked, etc.

Subcommittee members were encouraged to share permissions frameworks and deidentification rules used by their agencies to inform the work of these teams. If you would like to volunteer for a homework team, please contact LeAnn Fong-Batkin (lfongba@wested.org).

Finally, Kathy Booth noted that the December 3 joint meeting with the Common Identifier Subcommittee had been canceled. Given an independent assessment by Alvarez & Marsal (A&M), which indicated the CHHS Data Hub could not easily be scaled to support the needs of the Cradle-to-Career

Data System, future work of the subcommittee will focus on developing additional technical specifications for the data system.

Volume of Records

Erin Carter of WestEd reviewed responses to the survey about the volume of records for the P20W data set. While not all agencies had been able to respond, data from CCC, CDE, CSU, UC and apprenticeship totaled nearly 410 million records for the time period of 2008-09 to the current year. Records counts for the 2018-19 year were about 35 million.

The group discussed dimensions for data volume including the number of individuals in the system, the number of years of data, the number of data points about each individual, and the number of times the information will be stored. One community participant suggested examining the number of bytes of data being stored. Bruce Yonehiro from CDE echoed this idea and noted that the decision to not include all data points could be justified by comparing the proposed scale of the P20W data set to the full size of the data systems maintained by each data provider.

Subash D'Souza of CSU shared that in the development of his agency's data lake, they had examined the size of current data sets, starting with the number of students across the system. One lesson they learned is that they needed to access backup systems in order to include older records. He also noted that storing data is cheap—costs are incurred at the point that data are transformed.

Rodney Okamoto of CDE inquired what the starting year of the data would be. Kathy Booth of WestEd explained that early learning and care data are available as early as 2004-05. Alan Nakahara of CDE clarified that the first year of CALPADS data for K-12 students is 2008-09. Rodney Okamoto asked whether there will be a limit on how many years of data the Cradle-to-Career Data System will be allowed to store. For example, CALPADS is intended to store 20 years of data. Baron Rodriguez of WestEd noted that per FERPA, an end date is needed for data sharing, but that this date could be periodically reviewed and extended.

Hooman Pejman of UC asked whether estimates should include students who applied (whether or not they enrolled), as well as international students. Adding these records would increase the numbers for UC. Kathy Booth noted that the P20W data points approved by the workgroup include application data. Baron Rodriguez of WestEd added that the various committees had contemplated not including international students, particularly given concerns about the "right to be forgotten" that is part of European privacy laws. However, the Legal Subcommittee had recommended simply tagging international students to make the process of removing individual student's data easier so that international students could be included in the data set. Bruce Yonehiro explained that a subsequent decision to create an opt-out policy, which could be used by students regardless of their country of origin, removed the need for tagging international students for privacy purposes.

UC San Diego Super Computer Center Secure Data Enclave

Sandeep Chandra of the UC San Diego Super Computer Center described the work his organization has undertaken to create a secure data enclave, called Sherlock. Subcommittee members asked questions including:

What is the cost to participate in the data enclave?

Services are provided at cost. Rates depend on the scope of the data and the technical resources associated with implementation. The center tends to build each instance to the specific requirements of the client.

In cases where you offer multi-cloud storage, how do you create a single source of truth?

Clients normally use only one environment or establish the mechanism for weighting one data source more highly than another.

Some researchers make requests that require significant computation power, which costs more in the cloud. How do you contain these costs?

You can monitor usage and establish a threshold in cloud platforms. You can also tag the users or projects to identify who is accounting for the compute costs and charge them for those costs.

Solution Infrastructure

Kathy Booth of WestEd summarized A&M's framework for core components of P20W data systems and how these compare to the requirements set out in the Master Data Management Request for Information and the Intake & Request Processing Model (sometimes referred to as the "purple surfboard graphic"). A&M noted that the specifications did not address how data will be loaded, reviewed, and certified data; reference documentation will be managed; and data and tool usage will be tracked and evaluated.

Baron Rodriguez of WestEd reviewed an updated version of the Intake & Request Processing Model based on input from the Architecture Homework Team. Rodney Okamoto of CDE asked how data would be reviewed before they are released from the secure enclave. Kathy Booth clarified that data sets would be reviewed by the data provider. The managing entity would also maintain a small group of experts who can also help to examine data sets as needed.

The subcommittee then broke into small groups to identify the core technical components and staffing needed for specific functions of the data system, with the goal of identifying how costs could be determined. The section below reflects the small group discussions and information shared in the report out.

One strong theme across all groups is that it is very difficult to establish a specific budget at this point, given that several processes need to be nailed down. Costs will vary based on who is responsible for core activities, as well as which solution is selected. Some resources may be needed for the agencies to cover compute costs and for staff time to implement critical tasks associated with loading data into the system.

Two other concepts raised across the groups include: the value of leveraging state volume pricing structures, and the preference for data providers to use their own tools rather than adopting tools selected by the managing entity.

Finally, subcommittee members recommended surveying data providers to get a better understanding of their current capabilities and tools, as well as the estimated number of licensed users who would be using each type of tool (including documenting the number per role, such as entering, reviewing, approving, or modifying information).

Loading Data

The group noted that a number of items need to be further defined before costs can be determined, such as:

- What structure does the information have to be in when it is sent to the managing entity?
    - Will partners be able to load information using their existing file structures, which may include several different files (such as one for applicants and one for enrollments), or do all records need to be combined so there is one record per individual?
    - Will there be standard identifiers required for individuals?
    - If data must be transformed before it is loaded, what reference data needs to be included? For example, will each provider need to provide a specific tag for concepts such as gender?
    - Who will tag the data, the data providers or the managing entity?
    - What formats will be allowed for data uploads? For example, would .csv and .txt files be allowed?
    - Which cloud environment would be used, given that the type of standard format may vary by vendor?
- What type of ETL will be used?
    - Can data providers use their own ETL tools (for example, CDE is moving to Azure Data Factory but CSU uses Amazon)?
    - What will be expected of data providers that do not have ETL tools?
- Who will certify the data?
- How will further transformation be handled?

Visualizing Data

The group identified several lessons learned from implementing dashboards and query builders, including:

- Each question for the dashboards should be handled in a separate visualization, rather than trying to show all information on one dashboard
- Separate data marts may be needed for the dashboards and for the query builder, given that they will function differently
- The information for the dashboards should be cubed in advance to allow for instantaneous load times
- Provide an index that makes clear what types of data are available
- Include visualization tools that show information based on geography (like ARC GIS)
- Assume that backend maintenance will need to be done more than once a year to address security and software updates
- Budget with the assumption of larger up-front costs to nail down data definitions and to develop documentation that ensures users understand what information they are seeing. This will require time from program staff who are experts in the visualized topic areas, as well as engagement with users.
- Hire staff with the skills to do the visualizations and who understand the underlying analytics. This often means two types of positions: business analysts and data analysts.
- Staffing costs are likely to be greater than technology costs

Provisioning Data/Secure Data Enclave

The group identified several core components, including:

- Multifactor authentication for users
- Single sign on
- Statistical analysis tools
- Ticketing system to keep track of requests and approvals

The group noted the more people use the system, the lower costs become. In addition, custom tools drive up costs.

They also wondered whether it would be possible to pass the costs of computation in the secure data enclave to the users. Some small group members suggested establishing limits on transaction sizes or constraining the number of analytical tools (such as R, SPSS, Stata, Python, or SAS) available in the enclave.

Finally, the group recommended focusing on several specific categories of costs for the secure data enclave including labor, hardware, and software/licensing. The presentation from the UC San Diego Super Computer Center had estimated these costs at $2 million/year, with most costs being personnel. However, some small group members thought the costs might be higher, depending on the software provided and the licensing arrangements.