Cradle-to-Career Governing Board Staff Report

Date Report Issued:	October 25, 2023
Attention:	Members of Cradle-to-Career Governing Board
Subject:	**Update on the Data Infrastructure**
Staff Contact:	Dan Lamoree, Director of Data Infrastructure

At this meeting, the Governing Board will have the opportunity to learn about the progress of building the Data System.

Requested Action:

There is no requested action for this item. This is an informational item only.

Release 1.0 - Data Ingestion:

The timeline for the Data Ingestion (Release 1.0) milestone was highly ambitious. The target date of 2023-09-29 gave the team less than three months to deliver Release 1.0. Cradle-to-Career (C2C) is elated to report that despite the challenges associated with such a short timeline, the milestone was achieved. This was in no small part due to the collaborative partnership of the Office of Cradle-to-Career Data (Office), Deloitte, Strike Teams, Data Providers, Oversight, IV&V, and contractors.

Release 1.0 sounds relatively easy at first glance:

- Create Files
- Upload Files
- Ingest Files

Each of these "easy" lifts had to be scaffolded with the necessary planning, designing, developing, deploying, securing, debugging, documenting, and finally testing. Moreover, this work is then multiplied by our four environments (dev, test, stage, prod), which quickly compounds the total effort needed to

accomplish the milestone. All this, while orchestrating and collaborating with all the aforementioned partners.

Create Files:

Creating the data files was no small feat. The Data Providers worked with our internal C2C infrastructure and programs teams to construct File Specification documents. These documents have been mentioned in our previous communications, but the importance of these documents cannot be overstated; they are used to capture the data elements of the Participation Agreement (PA) mapped to the data as it exists in the originating Data Provider systems. Often no clear 1:1 mapping of PA element to source element exists, and this reality complicates the work of creating a unified P20W data set. The process of establishing the File Specification documents took months of work and updates to the File Specification had occurred up until recently.

Upload Files:

For the Data Providers to upload the Data Files, C2C needed to build a secure method for transmission and storage of the files. To do that, C2C leveraged cloud architecture that also provided for data protections in transit, at rest, and in use.

Once the file hits the designated server, a number of sequenced processes tigger. Processes like:

- scan files for malware;
- validate file structure;
- validate file payload completeness; and
- collect and store metadata around submission

Ingest Files:

Once the files are uploaded and validated, the data are then pulled into the data warehouse. While the complexity of moving from storage repository to data warehouse is straightforward, the data must then conform to the previously mentioned File Specification documents. These documents must be translated

from static documents to scripts used to create the necessary objects in the data warehouse. Each of these File Specification documents may require multiple domains (data files), each of which must be created in the data warehouse, along with the required tables for referential integrity. With this enforcement of referential integrity, submitted data are valid. Additionally, each element has to be categorized and tagged so the appropriate roles can be granted to users that govern access, masking policies, encryption, and so on.

Scaffolding:

Each of the high three level features of Release 1.0 must be scaffolded. The following includes some of the necessary support or processes that must:

- Multiple Environments
- Code Promotion
- CI/CD Pipeline and Deployment
- Git Code Repository
- Security

Multiple Environments:

As mentioned above, the Data System has four tiered environments, starting with Development, then Test, then Stage, and finally Production. This separation of environments is essential to ensuring that production data stays in production and only in production. Additionally, multiple environments reduces risks such as: reduced or eliminated downtown, thorough testing without impacting production, development without affecting production, and so on. As development progresses, and tests pass, the code used to build the system necessarily increases, but this happens in an orderly fashion using code promotion.

Code Promotion:

Given our multi-environment landscape, code is developed at the lowest level environment (dev) and then promoted to the next highest environment when all the necessary tests have been passed, no defects persist, and each line reviewed. This requires maintaining a git-based source code repository

integrated with project management tooling (and provisioning, access control, etc). The code repository is also integrated with our Continuous Integration / Continuous Deployment (CI/CD) pipelines, and allows for deployments, review, and versioning in a way that is transparent, easy to manage, and prevents the accumulation of technical debt.

CI/CD Pipelines:

These pipelines and methodology of code promotion from one environment to the next prevents schema drift. This helps keep the environments in sync with each other as new development is integrated into the Data System. Chiefly among the myriad of benefits of implementing CI/CD pipelines is the ability to quality assurance and risk reduction.

Security:

Finally, to make certain no shortcuts were taken for this incredibly tight schedule, C2C provided the Security Policies Task Force all the necessary security artifacts for review prior to accepting data into the Data System. This Task Force membership is made up of information security officers from each of the Data Providers, and meets once every other week. Some of these documents reviewed included:

- Privacy Threshold Assessment & Privacy Impact Assessment;
- System Security Plan (including NIST 800-53 Rev5 control families); and
- IV&V Finding Analysis Report.

These documents constitute hundreds of pages of content and give assurances to our Data Providers that their data are safeguarded.

Teams:

As alluded to above, the decision to procure multiple vendors to deliver the Data System was necessary. However, more vendors increases the necessary contract management, resource management, capacity planning, sprint management, and so on. To achieve this, a great deal of work was expended

up front because it was absolutely vital to the success of the project, despite compressing the schedule. Some of these activities included:

- sprint ceremonies
- approval cycles
- requirements refinement
- user story creation
- product backlog refinement
- sprint construction

In the end, the teams were able to coalesce into a cohesive development team and delivered Release 1.0.

Release 1.0 - Data Validation:

Our next release is a minor release, Release 1.1 which focuses on post-release stabilization, optimization, and automated data validation. The release methodology for the Data System uses a major-minor release pattern whereby major releases add major functionality to the system, increasing the security boundary or use of data. A minor release does not increase the security boundary but may add functionality of existing systems.

The bulk of Release 1.1 is to automate the validation checks made on ingested data, and the resulting responses to Data Providers. E.g., primary keys, foreign keys, and check table constraints. This validation ensures the data conform to the File Specification document, but also ensures the accuracy of the data to feed our master data management application for identity resolution. Identity resolution is our next major milestone whereby the Data System will be able to use machine learning to train a model that automates the process of probabilistically identifying individuals across the many existing disparate data sets. This process allows the Data System to truly act as an effective longitudinal system, linking students using existing data.

Gratitude:

The Data System is in excellent shape due to the engagement of hundreds of stakeholders and years of planning and discussion. With that said, the following individuals are recognized for their herculean contributions:

- Munny Chitneni, Deputy Project Director
- Mike Arakji, Chief Information Security Officer