Cradle-to-Career Data and Tools Advisory Board Staff Report

Date Report Issued:	February 26, 2024
Attention:	Members of Data and Tools Advisory Board
Subject:	**Intricacies of Data Ingestion and Breaking Down Data Silos**
Staff Contact:	Dan Lamoree, Director of Data Infrastructure

At this meeting, the Data and Tools Advisory Board (DTAB) will have the opportunity to hear updates about the development of the Data System.

Requested Action:

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

Release 1.1: Data Validation:

The release milestones for the Data System continue to be highly ambitious, and this release is no different. The target date for this release is March 2024. Cradle-to-Career (C2C) is pleased to report that despite the challenges associated with such a short timeline, we are on track to achieve this milestone. Release 1.1 extends the functionality of the Data System with the following features:

- Data Validation
- Tamr Infrastructure

Similar to Release 1.0, each of these deployments must go through the processes described in the November Governing Board meeting and detailed in the staff report. That is to say, they must be promoted through the environments, with extensive integration and security testing before each environment promotion.

Validate Submitted Data From Data Providers:

The data submitted by Data Providers goes through rigorous validation to ensure high quality data. This is vitally important when matching data between Data

Providers to produce the resulting P20W data set. The process for validation can be categorized into three buckets:

- Structure Validation
- Data Validation
- Data Profiling

Structure Validation:

Data Providers are required to upload their data files using the RFC 4180 CSV (Comma Separated Value) specification. This is an industry standard method for data transfer, and includes requirements such as:

- Delimiter (comma)
- Qualifier (double-quote)
- Terminator (carriage return and line feed)

The structure of the data, given the above specification, is then validated when uploaded to the Data System. Additionally, the first line of the data file must include a "header" which contains labels that describe each variable in the data. This header is used to check that the fields submitted in the file match the columns in the database, both in terms of naming and ordinality. These validations are necessary to ensure that the data file can be loaded properly into the data warehouse.

Data Validation:

Once the validation of the file structure is passed, the data values are validated, and include the following:

- Data Type
- Referential Integrity
- Regular Expression
- Uniqueness

The first check executed is to verify the data type can be converted from the character format coming from the CSV file to the appropriate data type in the data warehouse table, as defined in the file specification document. For

example, integer values for financial aid can be converted from the UTF-8 character encoding of the CSV file to the integer data type as it exists in the data warehouse table. Additionally, the data submitted must be within a specified length. For example, the first name of a student may be up to fifty characters, and a value exceeding this limit would be invalid. Lastly, data elements may or may not be nullable. Or, rather, the data column may require a value to be present for that element in a record. If a value is required and the submitted record does not have a value, the record is invalid.

Referential Integrity refers to the data adhering to a reference in a corresponding table. For example, the CDS (county-district-school) field in the California Department of Education (CDE) file specification is used to identify the school, and values submitted must reference a valid code maintained in a reference table.

Regular expressions are used to identify a pattern to match against. Some fields submitted may not have a reference value to enforce integrity. Instead they have a range of allowed values for the field in question. For example, the first name of a student may be up to fifty alphabetic characters, not to include any other character types (e.g., numeric).

Uniqueness refers to the fields of a record that constitute a distinct record. This collection of one or more columns defines the primary key of the table. By enforcing this constraint on the table, no duplicate records can be stored in the table. Essentially, this is how to identify a record within a table and join it to other tables sharing a similar constraint. For example, an enrollment table may store many student records, and those records would be unique to a student, by academic year, and school.

Data Profiling:

Data profiling is critical to understanding the accuracy and consistency of the data beyond what is considered valid. That is to say, the records may have

passed all of the validation checks, but then the data profiling reveals information suggesting there was an underlying problem in the data transfer. For example, if the gender variable for a statewide student demographics table only contained female values, the data profiling work would reveal that and highlight a problem in the underlying data transmission that the Office would troubleshoot with the Data Provider. Similarly, if the data profiling reveals a sudden unexplained shift in data trends over time, that could identify a problem in the data transmission. For example, if the enrollment rate of a school were to drop 90% from the prior year, then more investigation would be needed to identify the root cause.

Testing Synthetic Data:

No development is done using production data, and no production data traverse the production environment boundary. This means that the development of the Data System requires synthetic data to be generated for testing all the validations before promoting the code to the next environment. Synthetic data is data that is artificially created to take the place of real-world data. As part of testing this development, the team must execute both positive and negative test cases. Positive testing is defined by providing a valid input and returning an expected result. Essentially, data submitted must pass the validation checks described above. However, negative testing requires generating intentionally invalid data to ensure an error is returned.

Tamr Infrastructure:

Creating a comprehensive longitudinal data set from disparate source systems, that have undergone their own changes, updates and revisions, presents many challenges. One of these challenges is that no singular identifier can be leveraged to identify individuals as they traverse through the system. For example, CDE does not receive critical data elements from local educational agencies that are necessary for the Employment Development Department (EDD) to match and return the work history of an individual. However, some of these elements do exist in other data sets submitted to the C2C Data System.

Therefore, the data must be linked and unified to create a truly comprehensive longitudinal data set. That is to say, individuals must be matched and linked both within, and between, the data submitted by the Data Providers. To do this, the records must be consolidated to resolve conflicts in the data, and to create a singular source of truth. This process is commonly referred to as data mastering or golden record creation. The data validation and profiling steps detailed above are necessary to ensure the C2C Data System contains high quality data that feeds into the master data management solution, Tamr, tasked with matching and linking the data. Tamr leverages cloud infrastructure that efficiently and effectively scales to the volume of data, both in terms of number of records and size of records. This cloud deployment of Tamr infrastructure is scaffolded following the pattern of promotion and deployment described in the previous data infrastructure Governing Board staff report.