Data Challenge: Standardized Test Score File Processing

Overview

Duration: Approximately 60 minutes

Arrive ready to spend the first 5 to 10 minutes of the session presenting your solution. Explain the assumptions and important design decisions that you made.

The reviewers will then ask questions and explore particular areas more deeply. They will also propose new requirements. For these, we are looking to see how you incorporate them into the deliverable and how they might be implemented.

This audition will provide us an opportunity to see how well you problem solve, apply technical skills in solution design, communicate across audiences, and extend or modify an existing system.

- Use of the whiteboard (one will be provided) is strongly encouraged.
- You may distribute hard copies of any collateral, although lengthy presentations such as PowerPoint decks should be avoided.
- There are no restrictions on technology stack. Please employ programming languages, database platforms, supporting frameworks, and design approaches with which you are comfortable. You should be prepared to explain manipulations and write additional pseudo or actual code / queries.

Instructions

Using the following scenario, process the “sample-mcas.csv” file once so it aligns with the data format found in “sample-mcas-processed.csv”. Then modify your workflow to support an imagined volume of 20 similar files a day. How would you change the processing to accommodate this rate (20 files a day)? What about for 100 files a day? 1000? Design any procedures and/or scripts, as well as any required data model.

Please submit your proposed work flow and/or scripts and any required date model.

Scenario

The Ellevation platform can automatically bulk import student data from a school district’s School Information System (SIS), an extensive operational database. This allows a teacher to view a student record in the Ellevation platform without having to manually enter information already housed in the SIS.

In this scenario, Ellevation needs to process annual assessment scores for a client. “sample-mcas.csv” is an MCAS CSV file in the source format determined by the state of Massachusetts. Before our application can import the data, it needs to be manipulated in accordance with our internal canonical template.

“sample-mcas-processed.csv” is an example of a transformed output file and there are further specifications on the next page. For this scenario, the only pertinent tests are MCAS ELA, MCAS Math, and MCAS Science.

Specifications

NCESID: District unique identifier; Default to 373737

StudentLocalID: Not found in the MCAS format; Default to missing

StudentTestID: SASID field of the MCAS format

StudentGradeLevel: Stugrade field of the MCAS format; Grade level when student originally took the test

TestDate: Date test was administered; Not found in the MCAS format; Default to April 1 for MCAS ELA; Default to May 1 for MCAS Math; Default to June 1 for MCAS Science

TestName: Test cluster; Default to MCAS

TestTypeName: Test name within cluster; Default to MCAS ELA, MCAS Math, or MCAS Science

> In the source file, each record associates a student with 1 to 3 test results. For the processed file, each record should associate a student to exactly one test result. In other words, for each student, there should be a record for each existing test result (MCAS ELA, MCAS Math, MCAS Science).

TestSubjectName: Test subject associated with name; Default to ELA, Math, or Science

TestGradeLevel: Grade level for which the test was intended; Default to Stugrade field of the MCAS format

ScoreLabel: Score type associated with each subject; Default to Performance Level, Scaled Score, or CPI

ScoreType: Default to Level for Performance Level; Default to Scale for Scaled Score; Default to Scale for CPI

ScoreValue: Value of a score type

> Performance Level values can be found in fields “eperf2”, “mperf2”, and “sperf2” for ELA, Math, and Science, respectively.
>
> Scaled Score values can be found in “escaleds”, “mscaleds”, and “sscaleds” for ELA, Math, and Science, respectively.
>
> CPI values can be found in “ecpi”, “mcpi”, and “scpi” for ELA, Math, and Science, respectively.
>
> CPI and Scaled values can be brought in directly, but Performance Level values need to be mapped according to the code reference below.

F	>>	1-F
W	>>	2-W
NI	>>	3-NI
P	>>	4-P
A	>>	5-A
P+	>>	6-P+