CCER
Community Center
for Education Results

Data Engineer Assessment

Intro

Thank you for taking the time to interview with CCER. For the next step in the process, we want you to showcase your technical and design skills.

Below, you will find two publicly available files from the State of Washington about student enrollments and the demographics of teachers. We have also attached a school list from our internal CCER data.

Your task is to design and code a quick data pipeline to help answer questions about how closely the demographics of the educators and students align with schools in the Road Map Project region.

The best way to share this would be as a Github repository with in-line comments and markdown files, but you can use any version control or directory method you like. You are free to use any libraries or packages you are most comfortable with, we will provide some recommendations within this prompt that you utilize or ignore.

In an increasingly remote work environment, asynchronous communication ability and documentation are increasingly important, so we want to focus on your ability to communicate your thoughts as much if not more than your coding ability. Be sure to document your process, especially when you feel stuck or are running out of time.

Please do not spend more than 5-7 hours on this assessment.

Attachments

Utilize the below datasets for this project:

1. 2020-2021 Student Enrollment

2. 2020-2021 Teacher Demographics

3. Road Map Project School list (attached) - a tab-delimited file that includes the school code and district codes for all schools in the Road Map Region, which is comprised of the following school districts:

    - South Seattle (southern portion of schools in Seattle Public Schools district)
    - Highline Public Schools
    - Federal Way Public Schools
    - Renton School District
    - Auburn School District
    - Kent School District
    - Tukwila School District

Recommended Tooling

Below are tools we recommend for each step of the assessment. Again, if you feel like you are running out of time and cannot cover a specific area, just document your approach.

- Data Ingestion / munging: Python - requests, csv, json, pandas, etc.
- Data transformation: Python or SQL, leveraging Apache spark or dbt
- Data orchestration: Airflow, Dagster, Prefect, crontab
- Version control: Github, Gitlab

Exercise

The exercise is broken up into three sections of tooling described above. We have provided some guidance on which language to use, but feel free to work with the tools you know best. Code up the solution by completing the tasks below.

Often for ingestion and orchestration, Python will be most appropriate. For data modeling and transformation, SQL is easiest. However, feel free to use whichever tools you are most comfortable with. There are tradeoffs to whichever direction you choose.

Data Ingestion

1. Scrape or integrate the datasets from the webpage and flat file sources
2. Import the datasets as tables into a SQL database on your local machine
    a. Consider a cloud-based database. Why or why not?
    b. Do you develop validations, unit tests, or integration tests on this ingestion program?
        i. Why or why not?
3. Standardize the schemas to simplify the data transformation process
    a. Pivot the enrollments table to create a separate row to store the information in each demographic column before importing into the database.

Data Orchestration

4. Provide a mechanism or interface for orchestrating the data pipeline. This could be simple or complex, but think about how you would handle these datasets updating regularly at the source or stakeholders escalating a need to ingest it on a regular basis.

Data Transformation

5. In step 3, you standardized the schemas. Did you use a data transformations tool to do this?
6. Begin creating data models that can help analysts begin leveraging this data to answer stakeholders' questions
    a. Develop a model that normalizes these source tables into relevant dimension tables
        i. Schools
        ii. Districts
        iii. Races/Ethnicities
        iv. Etc. (get creative!)

b. Please submit a diagram or picture summarizing your schema design from source data to the data-modeling layer (DML)

i. How do you standardize and separate your sources from your core data models?

7. Develop relevant data transformations that can answer the below questions:

a. For each school type (high school, middle school, elementary school), please show the top 5 schools with the highest percentage of each of the following groups of students:

i. Black students

ii. Native American students (listed in the source file as American Indian/ Alaskan Native)

iii. Latine student (listed in the source file as Hispanic/ Latino of any race(s)),

b. Create a schools dataset that shows the total number of teachers at each school

c. For each demographic category, please show the 5 elementary schools with the highest number of teachers.

i. Do you leverage the same schools dataset as before? How or why?

d. How well do the teacher demographics represent the student demographics for each school?

e. Which 10 schools have the highest proportion of teachers of color?

f. How does this compare with each of these schools' proportion of students of color?

Exploratory Data Analysis

In closing out the exercise, please document some of your thoughts on the below questions:

- What other data would be helpful to collect in order to expand insights on these questions?
- What additional data models should we build?
- What additional analyses would be valuable to do?
- Anything else that you would like to explore or think about?

-