Summary

Data for Symmetric Health Solutions' platform comes from dozens of sources. For medical devices, each item in our database has attributes from most of these sources. To know that these attributes are for the same entity, we need to align records from these different sources based on any identifiers we can and master this data.

This assessment will have you do this for two example data sets that we use in our product.

GUDID - This is the American, FDA database of medical devices.
MDALL - This is the Canadian, Health Canada database of medical devices.

We would like you to write code to:

1. Download the latest release of each dataset
2. Align the data between the two datasets. In particular, we want to:
    a. Understand the general overlap between GUDID and MDALL.
        i. Records likely in both.
        ii. Records likely only in GUDID.
        iii. Records likely only in MDALL.
    b. Compare item descriptions between GUDID and MDALL.
        i. For records likely in both, make it easy to compare item descriptions.
        ii. We are not currently interested in other fields.
3. Make the aligned datasets queryable in some form (e.g., a local db we can run SQL against, or a structured file you can load into a jupyter notebook) to make it easy to examine and answer questions around 2a and 2b.

We have no preference as to which languages, libraries, or tools you use to accomplish this. We do, however, expect to be able to run your solution ourselves. So please either include instructions on installing anything we may need to run it, or better yet containerize the tools so we don't have to install anything new. In either case, please have a README to explain how to run. To give us access to the code, please create a public repo on Gitlab (or Github, or similar) and push your code there. Lastly, please spend **at most 4 hours** on this assessment. While there are endless improvements that could be made in the above process (caching, advanced techniques for better alignment, optimizing query speed, code quality, etc.) we are only looking for a very basic process that would be the start of something more advanced given more time. Feel free to include any comments on design decisions or future direction.

For context, one of our engineers spent ~2 hours to complete an MVP for the above tasks, with the majority of that time spent on the alignment portion. That leaves an hour or two for further enhancements. This context isn't meant as a comparison, but rather as a way to help you budget your time as you work on the assessment.

Details

The Datasets

The datasets are available at:

- GUDID:
    - There are different formats available for download, choose one.
        - XML: https://accessgudid.nlm.nih.gov/download
            - XSD schema file - https://accessgudid.nlm.nih.gov/download/schema
        - Pipe-delimited: https://accessgudid.nlm.nih.gov/download/delimited
            - Note that the pipe-delimited files are normalized, but the file names are not straightforward
    - The description is available in the **deviceDescription** field
    - The catalog number is available in either the **versionModelNumber** or **catalogNumber** field
    - The company is available in the **companyName** field
    - All of these fields are available in the pipe-delimited dataset in the **device.txt** file
- MDALL: https://open.canada.ca/data/en/dataset/c801a084-210b-4cd2-8513-26a00b66eb6f
    - Note that the “Access” links on this page are the datasets, but they open in the browser, and end up never loading there, so you would need to right-click and “Save as…” to download from the browser
    - These files are also normalized
    - The catalog number is available in the **Archived Device Identifier** and **Active Device Identifier** downloads under the **device_identifier** field
    - The description field is in the **Active Licence** and **Archived Licence** downloads under the **licence_name** field. These downloads can be joined to the identifier downloads on the **original_licence_no** field
    - The company name is available in the **Company** download under the **company_name** field. This download can be joined to the license download on the **company_id** field.

Alignment

For aligning the datasets, we suggest the following approach:

1. First, look for exact catalog number matches between the datasets.
2. Catalog numbers are not always entered in a uniform manner, so there may need to be some text cleaning to find exact matches

3. Since catalog numbers are not always unique, there will be cases where there are multiple items that may match. In these cases, other fields will need to be used to determine which catalog number match is the best
4. There may be inexact catalog number matches (e.g., a character at the beginning or end is missing) that should be matches

For this assessment, we would expect (1) to be working, but any additional steps (those above, or any you come up with on your own) would only be if you had extra time.

Output

At the end of the assessment, using whatever query tool is available, we would be able to expect to query for a particular catalog number and see the GUDID and MDALL descriptions. An example of an implementation might look as follows, but feel free to take any liberties you want to.

```
jboerner@jboerner:~/workspaces/assessment-demo$ time pipenv run search
Search for a catalog (blank for a random example, exit to exit): 1013470-180
Catalog number   : 1013470-180
MDALL Description: ARMADA 18 PTA CATHETER
GUDID Description: Armada 18 PTA Catheter 6.0 mm x 180 mm x 150 cm / Over-The-Wire
```

Besides trying a few catalog numbers like that a try, we would like some command to get the overall stats mentioned in 2(a). For example something like the following. The example is the minimum of what we expect. It would be great to see a bit more in-depth info here if you have time.

```
jboerner@jboerner:~/workspaces/assessment-demo$ time pipenv run overlap
gudid-only  catalog count: 1234
mdall-only  catalog count: 5678
catalogs in both datasets:  910
```