# icees-api

git: https://github.com/NCATS-Tangerine/icees-api

nginx:

ebcr.renci.org

v4:

run: ebcr0.edc.renci.org

/home/iceesuser/icees-api

is run under user iceesuser

url: https://icees.renci.org:16340/apidocs

data: /database/iceesdb/v4/data

v3:

run: ebcr0.edc.renci.org

/home/xuhao/icees-api

is run under user xuhao

url: https://icees.renci.org:16339/apidocs

data: /database/iceesdb/v3.\*/data

v1 and v2:

run: ebcr0.edc.renci.org

/home/xuhao/ddcr-api

is run under user xuhao

url: https://icees.renci.org/apidocs

data is stored in postgres natively under /database

smc:

run: ebcr0.edc.renci.org

/home/xuhao/icees-api{0,1,2}

is run under user xuhao

url: https://icees.renci.org:{16335,16336,16337}/apidocs

sample data from Kara

in case vm restart, need to remount /database. Otherwise, it is empty.

mount /dev/mapper/database to /database

# FHIR-PIT

git: https://github.com/NCATS-Tangerine/FHIR-PIT

How to run FHIR-PIT

### Setup

install sbt

Ensure running sbt from command line

install spark

Ensure running spark-submit from command line

install python >= 3.8

install dhall-json (https://github.com/dhall-lang/dhall-haskell/releases)

Ensure running dhall-to-yaml from command line

install virtualenv

Ensure running virtualenv from the command line

Create a virtual environment

Active virtual environment

pip install -r requirements.txt

Config

FHIR-PIT needs an input file that specify the pipeline. That file can reference other files. The files are written in the dhall format. Dhall is a configuration language that provides the flexibility for specifying FHIR-PIT pipelines.

Currently the files are organized as follows:

1. A pipeline.dhall which defined steps in the pipeline. This is the library

each step has a list of dependencies in the "dependsOn" field

the dependencies are given the conjunctive normal form

Inner bracket is an or. Only one needs to be met.

Outer bracket is an and. All must be met.

If can’t find python then update the python path is in pipeline.dhall (pyexec)

1. An example.dhall which defined which steps to run in the pipeline. This is the main config file

you can turn each step on/off by setting the skip to "skip", "reuse", "run"

"skip" will skip the step

"reuse" will reuse output of the step

"run" will run the step

when running the pipeline, it generates a file named "progress" that shows you the progress, a file names "report" is generated when pipeline ends it shows the summary. both files names can be configured in example.dhall

The paths to the dir requires by FHIR-PIT can also be configured in example.dhall

example.dhall requires the following dirs:

edit them in spark/config/example.dhall

1. config dir, default: /share/spark/hao/datatrans/spark/config (this dir should contain a file named icees\_features.yaml which contains the input to output mapping)

2. input dir, default: /var/fhir (dirs with name "FHIR" and "other" should be there)

3. temp dir, default: /var/fhir stores temporary data

4. output dir, default /share/spark/hao/data (dirs with name "icees2\_bins", "EPR\_binned" are created there by FHIR-pit) the data files are in EPR\_binned with filename pattern EPR\_binned<year><table>\_deidentified, the EPR\_binned\_bins.json and icees2\_bins/<year><table>\_bins.json are the bins

For the covid dataset

COVID/covid.dhall

1. config dir, default: /share/spark/hao/datatrans/spark/config/COVID

2. input dir, default: /var/fhir/COVID

3. temp dir, default: /share/spark/hao/data/COVID

4. output dir, default /share/spark/hao/data/COVID

### Build

Currently some tests are failing because of recent feature changes. Haven’t gotten to update the tests yet. They could be but are not necessarily bugs. We skip the tests:

cd spark

env SBT\_OPTS="-Xmx 64G" sbt "set test in assembly := {}" sbt assembly

### Run

Activate virtual environment

python spark/src/main/python/runPreprocPipeline.py spark/config/example.dhall

or

python spark/src/main/python/runPreprocPipeline.py spark/config/COVID/covid.dhall